Universiteit Leiden

Faculteit der Sociale Wetenschappen

Filtering a minority population among social media users

A case study on methodology in the field of big data

Gabriella Tisza

Master Thesis in Psychology
Methodology and Statistics Unit, Institute of Psychology
Faculty of Social and Behavioural Sciences, Leiden University
Date: 16 March 2018
Student number: s1968696
Supervisor: Julian D. Karch (internal), Marco Puts (external)

Acknowledgements

Foremost, I would like to thank my external supervisor, Marco Puts for the inspiring conversations through the course of this research work, and my internal supervisor, Julian D. Karch for his constructive criticism and challenging requests that helped me to improve the quality of my thesis.

I would like to thank Zsuzsa Bakk for the expert advice on the categorical data analysis and Piet J. H. Daas for that of social media.

Besides, my sincere thanks goes to Ali Hürriyetoğlu and Suzanne van der Doef for sharing their expertise in the field of text mining.

I thank my fellow interns, Chris Coemans, Jade Cock, and Guy Gerards for the great discourses that often gave me a different perspective.

Last, but not least, I owe my deepest gratitude to my husband, Adam B. Tisza for being a supportive and loving partner even in hard times, for taking care of my-, and our doggies' needs while my only focus was the computer, and for being a patient and thorough advisor of grammar and readability of my thesis.

Abstract

The aim of the research was to approach from the methodological point of view the question of filtering a hard-to-reach, minority population among social media users. For this purpose, a case study of the Dutch veg(etari)an Twitter population was used. Predictive performance was measured on three different data sets. These data sets reflect the main approaches for social media data collection, namely: choosing accounts arbitrarily and filtering their followers; analysing the social network of users; and analysing the tweets (text) of the social media users. For modelling, supervised learning techniques were applied. The results show that the highest predictive performance was reached when modelling based on social network data (F1-score: 0.90 – 0.97), whilst the lowest when on text (tweets) data (F1-score: 0.87 – 0.90). It is concluded, that filtering a minority population among social media users is possible with all of the three aforementioned data collection approaches. However, the highest predictive performance was reached when modelling on the social network data.

Table of Content

Filtering a minority population among social media users

A case study on methodology in the field of big data

Introduction

The aim of the research was to filter and predict a hard to reach minority population on social media. For this purpose the case study of the Dutch veg(etari)an[1] Twitter users was used. To reach this goal, knowledge from several scientific disciplines had to be gathered. Besides the methodological perspectives and the knowledge of statistical models, as it will be introduced, being familiar with the population of interest was of importance as well. Likewise, it required the understanding of the nature of social media data with the concerning social psychological aspects. Therefore the thesis begins with the introduction of these information assets.

The discipline of psychology has long been aiming for a better understanding of the human behaviour. Within the school of cognitive psychology, folk psychology provides a framework to map the human behaviour. *Folk psychology* (often referred to as *mindreading*) is a theory describing the human capability of explaining and predicting other's behaviour and mental state. According to the theory, the underlying dynamics of human behaviour are the same for every person, serving as a starting point for the understanding of others' intentions. This is where the term mindreading comes from. Understanding others' mind is the base for understanding their behaviour. Once one has an understanding of the other's intentions, the subsequent behaviour is considered to be predictable and explainable (Kashima, McKintyre & Clifford, 1998). This theory is closely related to the better known *theory of mind*, which describes the human ability to impute mental states to himself and others (Premack & Woodruff, 1978). However, while the theory of mind focuses on understanding the present state of mind, in folk psychology the accent is on the prediction. Therefore, folk psychology provides an appropriate theoretical basis for the current thesis. This theory allows to assume that a subpopulation - a homogeneous group of people - can be identified and predicted by their behaviour that originates from their interest, no matter on which platform. Applying this to the current research, it is assumed, that people's interests are reflected in their choice of

[1] For a better readability, the abbreviation of *veg*an* has been adopted, which is a well-known reference among internet users for vegans and vegetarians collectively without making a distinction.

users to be in contact with (i.e. social network) and their choice of topics to text about (i.e. tweets).

Population of interest

Although, a deeper understanding of the target population would not be crucial from a statistical point of view, the underlying theory requires, at least a basic understanding of the selected group, especially of their offline behaviour and driving motives.

In their qualitative study, Fox and Ward (2008) found that there are two main, ideologically different initial motivations: there are health (internally oriented) veg*ans and ethical (externally oriented) veg*ans. Health veg*ans are those who are motivated by personal reasons to sustain a healthier diet. It can be preventive to avoid sicknesses, or curative if already having some health issues. As for the ethical veg*ans, they are motivated by their concern for animals, hence their diet is fundamentally altruistic to forestall brutality against animals. Ethical veg*ans are linked to humanistic commitments whilst health veg*ans are linked rather to conservative and normative values. Moreover, regardless of the initial motivation, environmental and ecological concerns later contribute to the sustaining motivations. Hoffman, Stallings, Bessinger and Brooks (2013) add that ethical veg*ans are more convinced in their diet than health veg*ans, hence they often have more restrictions and commitments and follow the veg*an diet more persistently for a longer period of time. These conclusions were made based on the analysis of internet-based data, hence it demonstrates that “a large number of vegans and vegetarians, especially young, ethical vegetarians can be quickly and effectively reached using the Internet, particularly in social networking context.” (p. 142). Another, representative research scrutinised the beliefs of the Belgian population about vegetarianism and meat consumption (Mullee et al., 2017). According to the findings, 92.1% of the vegetarians agreed with the statement that meat production is bad for the environment. However, the agreement among meat eaters was only 19.8%.

In sum, the spectrum of the driving motives for following a veg*an diet is broad, however, the focus is on the concerns about animal welfare, health, and environment. Certainly, the clear differences between veg*ans and ‘others’ are that veg*ans do not consume meat, and regarding the reviewed literature, they seem to be more environment-conscious in general. Based on the theory of folk psychology, these characteristics are expected to be reflected in their social network and tweets. Hence, knowing the driving motives for following a veg*an diet has importance during the study.

Social psychological aspects

Since humans are social beings, they should not be studied without taking the social context into account. To do so, two phenomena are discussed here in details: *social desirability* and *deception*. Social desirability is the tendency of individuals to present themselves in a generally favourable way (Crowne & Marlowe, 1960). Deception is the act of intending to prompt a false belief or conclusion (Buller and Burgoon, 1996). They both can serve the higher need of being accepted by the group, hence belonging to it. Grossman (2017) explicates the effects of social desirability on social media users' online behaviour. *Social media* "are web-based services that allow individuals, communities, and organisations to collaborate, connect, interact, and build community by enabling them to create, co-create, modify, share, and engage with user-generated content that is easily accessible." (Sloan & Quan-Haase, 2017, p. 17.) Consequently, *social media users* are the people who make use of these services. The term *online behaviour* refers to the behaviour shown on social media platforms. As Grossmann concludes in her dissertation, social desirability is barely a motivation factor for deception. She refers to other authors, who have proven that the online deception rate is similar to that of regular conversations (Caspi & Gorsky, 2006). Moreover, it was also found that users report their opinion with a greater fidelity on social media platforms than in a formal context of, for example, an interview or survey (Cesare, Grant, & Nsoesie, 2017). Besides describing online behaviour, these findings indicate that when it comes to reliability, social media data can be comparable to conventional data sources (e.g. survey data, clinical interviews etc.). Therefore the reliability of the data sets used in this research and the universality of the research findings, based on these findings it shall not be questioned.

Social media

Undoubtedly, in the last decade, social media have become an integral part of most people's daily life in the developed countries, with a great influence on human behaviour, economics, and politics. It serves as a new platform for, among others, social life and interaction, information exchange, and dissemination of ideas and ideologies. In general, social media data belong to the group of *big data*. Big data are information assets that are too big and too complex to deal with in the traditional ways of data processing. The popularity of social media like Facebook, Instagram, and Twitter has resulted in unprecedented, continuously generated, huge amount of social data (just the Twitter messages count 456.000 per minute (!) worldwide[2]), which offers a new way to study human behaviour in general. Although social media data analysis is a novel field still under development, not only researchers and scientists show more and more interest in it, but the public sector as well (Cesare, Grant, & Nsoesie, 2017). Analysing social media data calls not only for novel methods to analyse big data, but also to process qualitative data (text mining) and even non-verbal data like emojis, pictures and sounds, not to mention the combination of these. When such a huge amount of data is available and generated from minute to minute, it is crucial to narrow the scope and be able to filter out the useful information.

The challenges of social media analysis are described the best by the six Vs: *volume, variety*, *velocity*, *veracity, virtue* and *value* (Williams, Burnap & Sloan, 2017). Volume refers to the amount of the exponentially increasing data production; variety to the multimodal nature of the data (text, images, audio and video); velocity to the speed of the data generation and of the response to the real world's events; veracity to the quality, accuracy, and reliability of the data; virtue to the ethics; and value to the added worth of understanding how the social world works.

In order to be able to study a population, especially in social sciences, the demographic properties (such as gender, age, income, education, marital status etc.) are usually of importance. They often serve as an initial point for defining the target population, or to compare or discriminate between two subpopulations (e.g. males and females, children and adults etc.). Most of the social media platforms, however, do not collect demographic information, making it difficult to identify the population of interest. While in the classic way of data collection, the demographic profile of the subjects is directly surveyed, in case of social media, this information has to be retrieved in a novel way. The demand for getting to know the users' demographic profile led to the increasing number of publications describing different approaches (see in Casera, Gant, & Nsoesi, 2017). Moreover, there are researchers who aim at understanding the composition of followers of Twitter pages. For example, Kavanaugh et al. (2012) investigated the Twitter page and news feed of 34 civil organisations' to better understand their audience.

[2] Source: https://globenewswire.com/news-release/2017/07/25/1058046/0/en/Domo-Releases-Annual-Data-Never-Sleeps-Infographic.html

Considering that social media data analysis is such a new field still in the exploration phase, it is acceptable, that it has no standards or general principles of model use yet. This is also reflected in the findings of Ruths and Pfeffer (2014), who found that large-scale human behaviour social media studies are generally poorly condcted, hence there is need for the introduction and implementation of higher methodological standards. After investigating scientific papers, they concluded that many of them misrepresent the entire population. Moreover, there are basic methodological issues, which could be easily prevented and/or corrected. The most important issues are related to a) population bias: when it is not taken into account that different kind of people use different kind of social media platforms; b) population mismatch: the proxy population often does not cover well the real population; c) improper use of machine learning techniques: e.g. if testing on the same data set where training of the model has been done, number of features are not taken into account when measuring performance; d) overfitting: when the analysis is overly tailored to the data, hence generalising the results is inappropriate; e) improper publishing of results; f) publication of only positive findings, which precludes the assessment of the extent of random chance.

Based on the introduced literature it is understandable that when approaching a social media research question, an interdisciplinary perspective is needed: “drawing on methodological traditions from across and outside of the social sciences, computer sciences and humanities” (Sloan & Quan-Haase, 2017, p. 7-8). This property of the social media offers an excellent base for a master thesis research for our interdisciplinary specialisation.

The thesis approaches social media data analysis from a novel perspective. Its main goal is to develop methodologically correct, generalisable ways to filter an arbitrarily chosen minority population among social media users (like homosexuals, vegans, artists, homeless people, victims of abuse, cancer patients etc.), who cannot be selected based on demographic properties. Despite Martinez et al. (2014) describe how they recruited members of a ‘hard-to-reach’ population (Spanish-speaking Latino gay couples living in New York City) for their study with the help of social media, their method cannot be generalised since it depends on connections with stakeholders of the target community, and involves webpage creations, meeting attendances and the training of the mentioned stakeholders. In the current thesis, possible means are demonstrated for the retrieval of a hard-to-reach minority population (Dutch veg*an population) from a social media platform (Twitter) while adhering to high methodological standards. The research is novel because its aim is to provide general, reproducible ways of data analysis for further studies. The main benefit of being able to filter such a subpopulation is that once it is retrieved, it can be further analysed. This way, subpopulations can be reached that otherwise would be extremely difficult to study by means of the classic data collection approaches (Golder & Macy, 2014).

Model performance

Since the composition of social media data is generally different from that of classic data collections, using the same reference values for the satisfactory level of a classifier's performance can be misleading. Therefore, the re-evaluation of the term of (expectable) predictive performance is needed. For this purpose other authors' work will be presented and used as reference. Chamberlain, Humby, and Deisenroth (2017) predicted 700 million Twitter users' age into three categories with 133,000 labelled data points based on what they followed (thus social network). Their overall Micro F1-score[3], used as main predictive performance measure, was 0.86. However, when further examining the results, the precision range was between 0.39 and 0.96, whilst the recall was between 0.50 and 0.95 across the age groups. Nguyen, Gravel, Trieschnigg, and Meder (2013) predicted the same age groups but with a different approach: they analysed text features (tweets), resulting in a Micro F1-score of 0.86 as well. However, their precision range was between 0.67 and 0.93 whilst the recall varied between 0.45 and 0.98 for the different age groups. Culotta, Ravi, and Cutler (2016) predicted Twitter users' demographics (gender, age, income, education, children, ethnicity, and political preference) based on their friends network, their tweet text, and the combination of these. Their F1-scores ranged from 0.56 to 0.87 with an average of 0.72 in the case of only the friends network analysis, 0.79 in case of only text analysis and 0.81 in the combined case. The predictive performance values of the aforementioned studies are summarised in Table 1 and regarded as reference values along the thesis. Chamberlain, Humby, and Deisenroth emphasise that analysing the social network is more time- and cost-efficient than analysing the tweets, although approximately the same accuracy can be achieved both ways. Their findings highlight that analysing only the tweets can result in a slight improvement compared with the social network analysis, but the highest accuracy is achieved when combining the two. Nevertheless, the improvement in the F1-score should be weighed against the extra time

[3] The Micro F1-score is the harmonic mean of the micro-average of precision and the micro-average of recall. In micro-average method the individual true positives, false positives and false negatives of the models are calculated, summed up, and applied to the statistics. For further information see Results section.

and costs required and against the fact that the text analysis is limited by specific linguistic and engineered features.

Table 1. Predictive performance reference values based on the introduced literature.

	Precision	**Recall**	**(micro) F1-score**
Network analysis	0.39-0.96	0.50-0.95	0.60-0.86
Text analysis	0.67-0.93	0.45-0.98	0.65-0.86

The current research aims to test the above-introduced techniques, namely the social network and the tweet text analyses, and compare the performance of the models. Nevertheless, instead of predicting demographic properties, the study attempts to find a minority population that cannot be filtered by demographic properties. This unique combination has not been researched and reported in the scientific world up until today.

Research questions

Based on the above-described theoretical background, the following research questions were determined: **(Q1)** Can the population of Dutch veg*an Twitter users be identified? If yes, **(Q2)** can the population be predicted based on their social network as well as[4] on their tweet texts?

Hypotheses

Based on the theory of folk psychology and previous Twitter data researches people's interests are presumably reflected in their choice of users to follow (social network, 'friends') and in their choice of topics to tweet about (word use, text data). It was seen as well that the different social media data collection approaches resulted in varied predictive performances. Hence, it was hypothesized that the population of Dutch veg*an Twitter users can be predicted based on their social network (**H1**) and tweets (text messages) (**H2**) as well. Moreover, it was expected, that the prediction accuracy, measured by the F1-score, will be better for the text data feature set than that of the social network (**H3**). However, regarding the cost-accuracy trade-off, it was expected, that the feature set of the network analysis will outperform those of the tweet analysis (**H4**).

[4] Approximately equal range of predictive performance measures.

Methods

Study design

The research described in this thesis had three main phases. Each phase included the analysis of a different set of data. The first data set was utilized to get the labels (veg*an or not) and a base social network model. The second was used for the extensive social network analysis, whilst the third for the tweet (text) analysis. Figure 1. depicts the main steps of the study, whilst details about the data and the processing methods can be found at the referring sections below.

Data

As the three data sets had different properties, they are presented one by one.

First data set. To obtain Dutch, veg*an Twitter users, first, accounts were pinpointed that were expected to be followed by Dutch veg*ans (e.g. pages of Dutch veg*an food brands, veg*an events, food blogs, and associations). In total, 24 accounts were chosen (see Appendix A) followed by 67,665 unique IDs. For more than 90% of the IDs, obtaining their metadata (see Appendix B) was not possible. These IDs were ruled out since the metadata were needed in order to label the users. 5,992 users remained as candidates for belonging to the target group of Dutch veg*ans.

The first step of the labelling intended to identify whether an account belongs to a Dutch. Manual labelling was used to maximise the reliability of the labelling. Semi-automatic methods were also considered but proved inapplicable. The manual labelling relied on the following heuristics: If the language setup was ‘nl’ and the time zone was ‘Amsterdam’, then the user was assumed to be Dutch. This way, 1,150 IDs were identified as Dutch. The remaining 4,842 accounts were manually labelled based on the collected metadata: the language setup, the profile location, and the content and language of the profile description. If any of these implicated that the user was not Dutch, then (s)he was ruled out. The Dutch label was given exclusively when the user could undoubtedly be identified. That was the case in 4,269 of the 5,992 users.

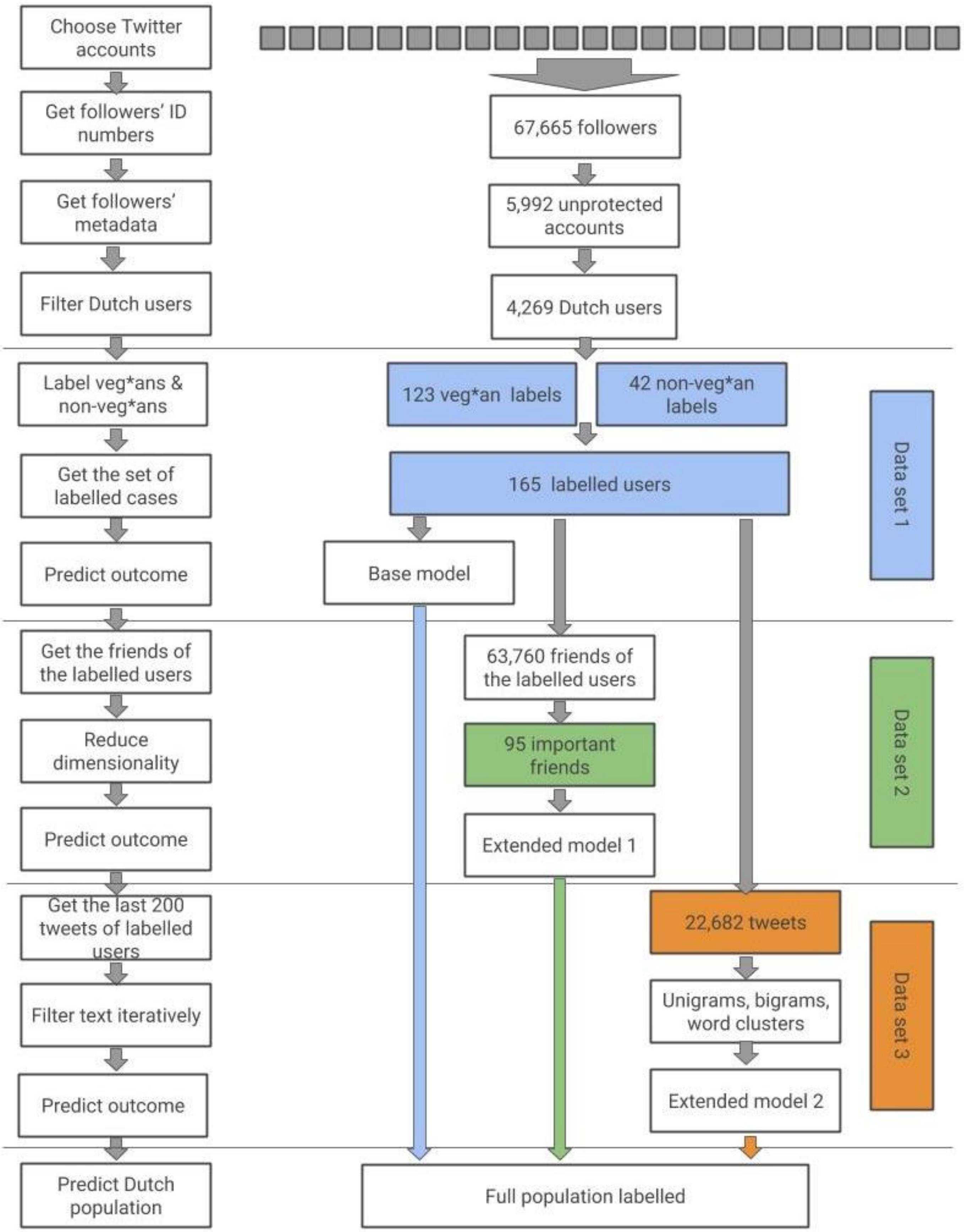

Figure 1. The study design

After separating the Dutch, the veg*ans had to be identified. To maximise the reliability, manual labelling was chosen again. If there was a clear indication in the profile description that the person was veg*an, then (s)he was classified so, which was the case with 123 users. The criterion of being non-veg*an was either a clear indication in the profile description (e.g. “husband, dad, meatlover”), or a (re)tweeting of recipes with meat. In total, 42 users were labelled as non-veg*an. (Since the population at hand had been collected via veg*an related accounts, veg*ans were expected to be overrepresented.)

The final data set – after the data cleaning – consisted of 165 users, one outcome variable, and 24 feature variables. The outcome variable marked whether the user was veg*an or not, whilst the feature variables referred to which of the 24 arbitrarily chosen accounts were followed by the user. Since multicollinearity was an issue at some of the chosen methods, features with a higher than 0.90 correlation coefficient were to be removed according to the recommendations of Tabachnick and Fidell (2012). As the data set did not contain highly correlated features, no feature needed to be removed. One of the accounts was not followed by any labelled users though, it was hence removed. Further steps of data pre-processing – such as scaling and standardizing – were not necessary due to the binary nature of the data. The outcome variable and the remaining 23 feature variables formed the basis for the data analysis.

During the analysis, if a model required certain hyperparameters, an automatic grid search was executed. Then, based on the rough estimation of the automatic grid search, the hyperparameters were manually refined and used for the model. Refining the hyperparameters was an extra step to maximise the predictive performance. Since the automatic grid search systematically probes random values for the hyperparameter optimisation, it uses values located to a given distance from each other. The manual refining targeted the gap (distance between the optimal and the next fitted value) around the optimal hyperparameter value, by forcing the model to ‘try’ those values as well. That is, the models were trained again with a five-times repeated, 10-fold cross-validation on the manually defined set of hyperparameter(s). In some of the cases this resulted in a differing optimal value, in other cases the optimal value remained the same as it was found by the automatic gird search.

Besides, the importance of the features was calculated by an in-built function (caret package, *varImp*): based on a sensitivity analysis that measured the effect on the output of a given model when the inputs are varied. It is a scaled metric, therefore it was suitable for comparison between the different models.

The hyperparameter optimisation and the feature importance calculation was done the same way on all of the data sets.

Second data set. The second data set consisted of the friends network of the 165 labelled users from the first data set. In this round, all of the friends' IDs was downloaded. From the original 165 cases, 147 users' data were possible to retrieve. From the 147 users, 108 were veg*ans and 38 were non-veg*ans. In total, a list of 87,780 ID numbers were downloaded, of which 63,760 were unique. The other 24,020 IDs were recurrences, i.e. IDs that appeared in the list multiple times due to being friends of multiple users.

Since the data consisted of a huge number of features (i.e. user IDs that they follow), the first step was to filter the most important ones – which also contributed to the reduction of noise and the chance of overfitting the data. To this end, the analysis of the repeated variables was conducted separately for the veg*an and the non-veg*an group. The most important features were selected by the following heuristic: if an ID number was observed at least in 25% of the group, then it was considered important hence kept. Therefore the most important predictors for the veg*an group were the ones that were followed by at least 27 people, which occurred in 21 cases (see Appendix D). Consequently, the most important predictors for the non-veg*an group were the ones followed by at least 10 people, which occurred in 78 cases (see Appendix E). After combining the features and removing the highly correlated ones, the data set was ready for the analysis of the 147 cases, 95 features (see Appendix F) and one outcome variable (veg*an or not). Further steps for the data pre-processing – such as scaling and standardizing – were not necessary as the data were binary. During the analysis, the manual refining of the optimal hyperparameter values and the variable importance was calculated in the above-described way.

Third data set. The third data set was downloaded on 16 November 2017, and it consisted of the recent 200 tweets of the labelled users including their metadata (e.g. date of creation, whether being retweeted, favourite counts, geotags, coordinates etc.). This time, the data of 145 users could be retrieved. Multiple options were considered regarding the number of tweets to download. Based on the findings of other researchers, the last 200 tweets provide sufficient information for reliable classification; analysing more data yield insignificant gain in performance (Sap et al., 2014; Volkova, van Durme, Yarowsky & Bachrach, 2015; cited in Morgan-Lopez, Kim, Chew & Ruddle, 2017). Therefore the decision was made to download the last 200 tweets per user, which resulted in 22,682 tweets in total as some users had less than 200 tweets. These tweets consisted in 137,931 unique words (tokens), from the time-range of 15-11-2009 and 16-11-2017.

Once all tweets had been collected, they were pre-processed and converted into numerical format. During the pre-processing, the tweets were cleaned from characters, URLs, and *stopwords*. Stopwords are expletives, that is, words belonging to the natural language, with little or no meaning (e.g. *a, an, the, and*). Since the tweets were retrieved from Dutch users, yet contained English text too, a unique list of stopwords was created (see Appendix H). This list included the default English and Dutch stopwords from the *quanteda* package (Benoit, 2017), and additions, which are very typical for online text, but again, have little importance (e.g. *via*). Then, the tweets were divided into tokens (simple words or word combinations used as features). Thereafter, the importance of each token was defined by a weighting scheme. For this purpose, the *TF-IDF* (term frequency-inverse document frequency) of the words was used, which is a numerical statistic, quantifying the importance of the words in a given text. The importance is measured based on the frequency by assigning weights to the words, whilst taking into account that some words occur more often than others. Based on the literature review of Beel, Gipp, Langer, and Breitinger (2015), who reviewed articles from 1998 to 2014, this method was the most popular weighting scheme.

Besides the text mining tools, *K-means clustering* was used for dimension reduction as well. It is an unsupervised classifier, which aims to partition the observations into coherent clusters by assigning the observations to the nearest mean whilst minimising the within-cluster variance.

Since the feature set had to be optimised, the text data pre-processing was done in an iterative way. Initially, the data were pre-processed, the models were fitted, and their performance was measured. If the model required certain hyperparameters, their optimal value was determined by the above-described way. Also, the variable importance was calculated as introduced before. Based on the output, the pre-processing was refined, and the models were fitted on the newly tailored data set. There were six filtering rounds in total, which were administered as follows:

In the first round, the data set consisted of single words (unigrams). The # and @ characters[5] were retained in front of the words, as they were considered to be organic parts of the text. However, the data was cleaned by removing other characters (including emojis), URLs, and default stopwords. Thereafter, the correlation among the features (words) was

[5] The # (hashtag) in the Twitter environment is used to categorise content and track topics, whilst @ (mention) is used to invite users into conversations.

computed and the highly correlated ones (r > 0.9) were removed. After the data cleaning, the data of 115 users remained to be relevant with 123 features.

In the second round, the same filtering setup was used as in the first round, except that the correlating features were not removed. This step was implemented to test the effect of the serious feature loss in the previous step. In this round, the data set of 115 users and 1321 features was used for the modelling.

In the third round, the stopword list was adapted to the data, that is, the negative words *no*, *not*, *nor*; *niet* (= not), *nee* (= no), *geen* (= none), *zonder* (= without), and *worden* (= become) were removed from the list whilst *via*, *http*, and *co* were added to it. This step was taken as veg*ans were expected to text about becoming veg*an, being environmental conscious, and sharing information about their diet (which could be related to food without animal based ingredients as meat, eggs, milk, etc.). The filtering left 116 relevant users with 1325 features (single words) in the data set.

In the fourth round, the # and @ were removed from the beginning of the words in order to test the importance of these characters. This step resulted in 123 relevant users with 1115 features (single words).

In the fifth round, bigrams (combination of two words) were generated from the data of the third filtering round (which later appeared to be the optimal filtering setup). Then, the most important ones (defined by setting the threshold of relative document frequency[6] to 0.98) were selected and used as features. Based on their offline behaviour, veg*ans were expected to use different word-combinations than others (e.g. *without meat*, *no eggs*, etc.). This resulted in a data set of 144 relevant users with 622 features.

In the sixth round, the unique words of the third filtering round (which later appeared to be the optimal filtering setup) were clustered with K-means clustering. The aim of the clustering was to reduce the dimensionality. The optimal number of clusters was decided upon based on the scree-plot, that is the within cluster variation – number of clusters trade-off was optimal, when the 1325 unique words were assigned to 120 coherent clusters. After the data cleaning and clustering process, the data of 116 users remained relevant with 120 clusters as features.

[6] The probability that a given document *d* contains a term *t*.

Tools

Data retrieval tools. For the data retrieval, two tools were used. One was the *Jupyter Notebook* (Kluyver et al., 2016), an open-source web application, which, among others, allows the use of the Python programming language. The other was the *Rstudio* (2016)**,** which is the open-source environment for the R programming language. They both require an interface that enables communication with the Twitter platform, thereby allowing data download. *Tweepy* (Roesslein, 2009) is the open-source Python library, which was used to connect to the Twitter Streaming APIs (application programming interfaces), and *TwitteR* (Gentry, 2015) is the corresponding R package. Furthermore, for the main analyses, the Rstudio 1.1.383 (2016) and the *caret* package (Kuhn, 2017) was used.

Classifiers. When considering the classifiers to apply, several aspects were taken into account. The data set was assumed to have a high dimensionality with a binary outcome, to be sparse, and probably imbalanced with correlated features. Moreover, at least in some of the data sets the data itself was expected to be binary (friend/follower or not). Both classic and cutting-edge classifiers were considered. The use of H2O[7] was planned, however, the slice of big data at hand did not require special interfaces for big data statistics. It was important to choose the classifiers so that they would be applicable to all data sets, thereby allowing for the comparison of their performance at the end of the research. Hence, only supervised techniques were selected for the modelling.

After having the criteria fixed, the range of the possible methods was mapped. As mentioned above, given that social media research is a new field of data analysis, there are no official principles or standards for model use. Therefore, related papers were studied, the assumptions and ways of applications of these methods were examined, the SAGE Handbook of Social Media Research Methods (Sloan & Quan-Haase, 2017) was scrutinized, and experts were consulted (just as Zsuzsa Bakk at the faculty – who has expertise in the categorical data analysis; and Ali Hürriyetoğlu at the Statistics Netherlands – who has expertise in the Twitter data analysis). After obtaining sufficient information, the most suitable methods were selected to analyse the data.

As declared above, it was important to use both classic and modern classifiers. From the long list of modern classifiers neural network, deep learning, and Support Vector Machines

[7] A leading open source platform, which makes it easy to apply AI and deep learning to solve complex, big data problems (Aiello, Eckstrand, Fu, Landry and Aboyoun, 2017).

were considered suitable for the presumed construct of the data. However, once the data were at hand, analysis with deep learning techniques were proven to be inapplicable due to the dimensionality of the data. The final choice of models are introduced in details below. The applied methods were selected from six classification approaches. They consisted of linear classifiers, tree-based methods, simple probabilistic classifiers, non-parametric classifiers, maximum margin classifiers, and Fuzzy-Rule Based Classifiers.

Penalised GLM Logistic Regression is one of the classic methods, which is often used along with machine learning techniques despite its simplicity. It models how a binary response variable depends on the explanatory variables (features) by maximizing the likelihood. The model estimates the probability of the binary response given the features. For increasing performance on high dimensional data, the regularisation parameter lambda was applied. The optimal lambda value was found by automatic grid search with five-times repeated 10-fold cross-validation. This method was chosen due to its ability to handle high dimensional data with binary outcome.

Boosted Tree with Adaptive Bosting is a method, which grows a sequence of simple, binary trees, where each succeeding tree is built on the prediction residuals of the previous tree, this way reducing the final residual variance (error). However, to prevent overfitting, the optimal number of trees needed to be defined, for which automatic grid search with five-times repeated 10-fold cross-validation was used. From the several available boosting techniques, the Adaptive Boosting was applied, because it uses stage weights for the model improvement. In this boosting technique, the weighting is based on the error on each given stage. Therefore, the process chooses only the most important features, which are known to improve predictive power, thereby reducing dimensionality. This property made the classifier a suitable candidate for analysing high-dimensional data.

Random Forest is an approach that grows a given number of classification trees, which then one by one classify the input vector. In the end of the process, the class label is given, which for came from the single trees the most ‘votes’. The important qualities of the Random Forest are that it efficiently handles large data sets without the need for feature deletion, it gives an estimate for the most important features, handles missing data efficiently, and balances error of imbalanced data sets. Moreover, it can be used on labelled and unlabelled data, it does not overfit the data, and mislabelled cases can be detected by using the outlier measure.

Naïve Bayes is one of the simplest classifiers among the machine learning techniques, which is based on the Bayes probability theorem. In comparison with the Logistic Regression, the Naïve Bayes optimises the joint probability (assumes conditional independence), whilst the Logistic Regression optimises the conditional probability. Despite being a simple model, it is demonstrated to compete in performance with the Support Vector Machines (Rennie, Shih, Teevan, & Karger, 2003). The Naïve Bayes classifier has no limitations regarding the number of features, and it is suitable for data with a binary outcome. Furthermore, although assuming strong independence between the features, it has been proven to perform well even with correlated data (Zhang, 2004).

Support Vector Machines are non-probabilistic, linear, two-class, maximum margin classifiers, which work with a separating hyperplane with a gap around it as wide as possible. The function of this separating hyperplane is to separate the points belonging to the different classes. The classification is done by detecting the extreme values (borders) of a class. There are several options for *kernel* choice, from which three were selected. (Kernels are similarity functions, which enable the machine to operate in a high-dimensional, and/or nonlinear feature space by enlarging the space by making the decision boundary linear.) There is a slight difference in performance between the SVMs with different kernels, depending on the type of feature-space. The *linear kernel* is widely used and known for high efficiency for linear problems, the *radial kernel* allows the modelling of strongly nonlinear and infinite dimensional problems, whilst the *polynomial kernel* is the best suitable for high-dimensional data. The performance of the classifiers highly depends on the hyperparameter selection, for which automatic grid-search was used. The optimal value of the hyperparameters was defined by five-times repeated 10-fold cross-validation. Moreover, the used R package (*e1071* (Meyer, Dimitriadou, Hornik, Weingessel,& Leisch, 2017) embedded in *caret* (Kuhn et al., 2017)) is suitable for sparse data matrix as input data.

Learning Vector Quantization is a non-parametric, prototype-based classification algorithm from the field of artificial neural networks with winner-takes-all (competitive) learning strategy. If comparing the LVQ with the SVM, LVQ is based on the class representatives (prototypes), which are class-typical sensitive vectors inside the class distribution area, whilst the support vectors are the extreme values (borders) of a class. It is a novel approach, which is a valuable alternative for SVMs (Kaden, Riedel, Hermann, & Villmann, 2014; Nova, & Estevez, 2014). The method is suitable for high-dimensional, sparse data, which makes it a reasonable choice for the data structure.

Fuzzy Rule-Based Classifier with Chi-algorithm is based on space partition approach, hence it is comparable to the SVM. Despite the fuzzy logic has been known since the 1960s, its use in the field of big data analysis has recently been discovered. For the analysis, the Chi algorithm was used, which has been referred to as one of the most popular, cutting-edge machine learning approaches for big data analysis (Elkano, Galar, Sanz & Bustince, 2017). Besides being popular, the fuzzy logic is proved to be an excellent supervised learning tool for classifying big data with binary meta-features (Kowsari et al., 2018). In fuzzy classification, an item can belong not only to one but to several different classes to different degrees. Then, the membership value is evaluated and the final classification is done by the algorithm. The membership values are percentages in the range of 0 and 1, summing up to 1.

Results

The results of the analyses are introduced across three performance indicators. The chosen measures, by their nature, take into account the slight imbalance in the data sets (Bekkar, Djemaa, & Alitouche, 2013). For the good comparability with previous researches, the F1-score, and due to the slight data imbalance, the balanced accuracy were chosen as predictive performance benchmarks. To measure general model fit, the McNemars's test value was assessed. All of these measures use terms that are related to the *confusion matrix*. The confusion matrix is a 2 x 2 table, which reports the number of true positive, false positive, true negative, and false negative cases (see Table 2). Whilst the true positive and true negative cases are the ones that are correctly classified, the false negative ones are the cases that were observed as positive but classified as negative. Similarly, the false positive cases are the ones that were observed as negative but were classified as positive.

Table 2. The confusion matrix and the terms behind.

		Observed	
		Negative	Positive
Predicted	Negative	True negative (TN)	False negative (FN)
	Positive	False positive (FP)	True positive (TP)

The performance measures in detail are as follows:

Balanced accuracy takes into account the accuracy[8] given the class, thereby correcting for the imbalance. Because of this added value, the balanced accuracy is preferred above the normal accuracy. The balanced accuracy is calculated in the following way:

[8] The percentage of the correct predictions from all of the predictions that have been made.

$$Balanced\ accuracy = 0.5 * \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

F1-score or balanced F-score is an overall measure of model accuracy. The F1 score is generally used in classification problems. It is the harmonic mean of the precision and the recall, computed as follows:

$$F1 - score = \frac{2 * precision * recall}{precision + recall}$$

Where the *recall* is the proportion of the correctly identified positive cases over the true positive plus the false negative cases:

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

And the *precision* is the proportion of the true positive instances over the true positive plus the false positive instances:

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

McNemar's test tests the *marginal homogeneity* of dichotomous data in a 2 x 2 contingency table. The marginal homogeneity determines whether the row and column marginal frequencies are equal. The null hypothesis (H_0) is that the marginal homogeneity is the same for each outcome. When the test value is not significant ($p > 0.05$), then the marginal homogeneity does not differ statistically. This means that the predicted values do not differ significantly from the observed values. In other words, the model fits the data.

First data set

The first data set contained the followers of 24 veg*an related accounts and their metadata. This data set aimed to assess whether it was possible to identify the Dutch veg*ans on Twitter (research question **Q1**). In addition, this data set was used to set up a base model for predicting veg*ans based on their social network. To this end, first, the data were randomly split into training and test sets while keeping the distribution of the outcome variable. Although the distribution of the outcome variable was not meaningful in this case, this step was taken to avoid unequal random sampling from the relatively small group of negative cases (non-veg*ans). If the distribution of the outcome variable at the training and test sets had been unequal, the insufficient number of negative cases could have caused low performance of the models. Furthermore, the models were always trained on the same 2/3 of the data while retaining the remaining 1/3 for testing the model accuracy.

For each model, a five-times repeated, 10-fold cross-validation was applied. When the model required certain hyperparameters, an automatic grid search was executed. Then, based on the rough estimation of the automatic grid search, the hyperparameters were manually refined and used for the model.

For modelling, the above-described nine techniques were applied: penalised GLM Logistic Regression, Fuzzy-Rule Based Classifier with Chi algorithm, Boosted Tree, Random Forest, Learning Vector Quantization, Naïve Bayes, and Support Vector Machine with radial, linear, and polynomial kernel. Since the Naïve Bayes classified every instances to the majority group, its prediction was considered unreliable and therefore was excluded from the presentation. Moreover, because the McNemar's test value was significant ($p < 0.05$) in case of the Fuzzy-Rule Based Classifier and the SVM with radial kernel, the performance of these classifiers was regarded as unreliable for the given data set.

As for the model performance (see Table 3), the balanced accuracy and the F1-score will be discussed. From the fitted nine models, the Boosted Tree resulted in the highest balanced accuracy (0.8087), whilst the Support Vector Machine with polynomial kernel in the lowest (0.6288). Regarding the F1-scores, the Boosted Tree provided the highest value again (0.9136), leaving the other models slightly behind. In sum, the F1-score value ranged between 0.8101 and 0.9136 across the different classifiers, whilst the balanced accuracy ranged between 0.6288 and 0.8087.

Regarding the features, on average, two associations for Dutch veg*ans (*NLvegan* and the *vegetariersbond*), a political party (*PartijvdDieren*), a foundation for plant-based diet (*vivalasvega*), a plant-based food brand (*vivavivera*), and a veg*an related news and food blog (*HeavenofDelight*) were among the five most important ones for the models (see Appendix C). Despite having ranked features, most of the models reached the highest accuracy when using the majority of the 23 provided features.

Table 3. The performance of the succeeding models on the first data set.

	Logistic Regression	**Boosted Tree**	**Random Forest**	**LVQ**	**SVM linear**	**SVM polynomial**
Balanced accuracy	0.7962	0.8087	0.7452	0.6308	0.7596	0.6288
F1-score	0.9000	0.9136	0.8750	0.8101	0.8219	0.8736

Second data set

The second data set, as an extension of the first data set, contained the whole social network data of the labelled users. It was used to assess the predictive performance based on the social network as feature set, thereby finding answer to the related research question (**Q2**: Can the population be predicted based on their social network as well as on their tweet texts?).

For the good comparability, the same methods and statistical models were applied to the second data set as to the first. First, models were trained on the training set, then the accuracy of the trained models was assessed on the test set. For each model, five-times repeated, 10-fold cross-validation was applied. When the model required certain hyperparameters, an automatic grid search was performed. Thereafter, based on the rough estimation of the automatic grid search, the hyperparameters were manually refined and used for the model. Although the subjects in the separated training and test sets were not necessarily in overlap with those in the first data set, the applied performance measures are standardised, therefore they were comparable with each other.

This time, the SVM with radial kernel did not assign anyone to the minority group, therefore this model is not introduced in details. Moreover, because the McNemar's test value was significant ($p < 0.05$) in case of the Fuzzy-rule based and the Naïve Bayes classifiers, the performance of these classifiers was identified as unreliable for the given data set.

Evaluating the models' predictive performance based on the second data set (see Table 4), the balanced accuracy ranged between 0.7774 and 0.8333 whilst the F1-score was between 0.9014 and 0.9667. Having an extended number of features compared with the first data set resulted in new users among the most important ones. This time, on average, multiple food blogs (*CulyNL*, *FoodiesMagazin*, and *LekkerTafelen*), a news and food blog (*FoodReporter*), a sustainable lifestyle blog (*boerenfluitjes*), and one of the Dutch veg*an associations (*NLvegan*) were among the five most important features (see Appendix G).

In sum, the performance of the fitted models was slightly better on the second than on the first data set.

Table 4. The performance of the succeeding models on the first and second data sets.

	Logistic Regression	Boosted Tree	Random Forest	LVQ	SVM linear	SVM polynomial
Balanced accuracy	0.8190	0.8333	0.7905	0.7774	0.8333	0.8190
F1-score	0.9315	0.9459	0.9014	0.9189	0.9667	0.9315

Third data set

The third data contained the tweets (text data) in a numerical format. This data set was used to assess the predictive performance based on the text data as feature set (**Q2**).

For the good comparability, the same statistical models and methods were applied at the third data set as at the previous ones. The earlier selected performance measures (balanced accuracy and F1-score) are compared between the different filtering setups on the third data set in Table 5. For more performance measures, see Appendix I-N.

In the first step (data set 3/I), the correlated variables were removed as previously. This resulted in a large feature loss (from 1321 to 255). Six out of nine classifiers were not able to perform this task, that is, they either assigned all instances to the majority class or simply broke down (see Appendix I). Only the Support Vector Machines had assessable performance, from which the SVM with radial kernel was identified as unreliable due to significant ($p < 0.05$) McNemar's test value. The balanced accuracy of the SVM with linear and polynomial kernel was 0.6865 and 0.6488 , whilst the F1-score was 0.8814 and 0.8852 respectively.

Next (data set 3/II.), to assess the effect of the feature loss in the previous step, no further features were removed after the data pre-processing. That is, the modelling was performed with the complete 1321 features. In this case, the Support Vector Machines, the LVQ, and the Random Forest resulted in assessable performance. From the SVMs, the SVM with radial and liner kernel were excluded from the reliable classifiers due to significant ($p < 0.05$) McNemar's test value (see Appendix J). The balanced accuracy ranged for the remaining classifiers between 0.6310 and 0.7798, whilst the F1-score varied between 0.8667 and 0.8929.

In the third step (data set 3/III.), the data were pre-processed with the tailored stopword list (see Appendix H) then the models were trained and tested. In this round, the penalised Logistic Regression, the Random Forest, and the SVM with polynomial kernel had assessable performance. The rest of the classifiers either failed to converge, predicted only instances belonging to the majority class, or had a significant McNemar's test value (see Appendix K). The balanced accuracy of the remained, reliably performing classifiers were between 0.6310 and 0.7242 and the F1-score fluctuated between 0.8667 and 0.9000.

In the fourth step, the data (data set 3/IV.) were analysed after removing the # and @ characters from the front of the words. In this round, five out of nine classifiers did not provide an assessable performance. All of the other four classifiers had a significant McNemar's value, hence they were not regarded as reliable despite producing an output (see

Appendix L). In sum, none of the nine classifiers were able to produce a reliable performance on this data set.

In the fifth round, the models were fitted to the bigram data (data set 3/V.). On this set of data, seven out of nine classifiers failed either by breaking down, not converging, or predicting only the majority class labels. Although the Logistic Regression and the Random Forest classifiers produced an assessable outcome, these were considered unreliable due to their significant McNemar's test value (see Appendix M). This means, that predicting from the bigrams resulted in an uninterpretable outcome. Therefore, the planned analysis of trigrams was discontinued.

In the sixth round, the models were fitted to the clustered data (data set 3/VI.). In this round, six out of nine classifiers did not produce an assessable performance either due to breaking down or assigning every instances to the majority group. The SVM with polynomial kernel was also excluded from the reliable analyses due to significant McNemar's test value. From the remaining two classifiers, the Boosted Tree had a very poor performance with a balanced accuracy of 0.5933. However, this classifier recognised the same five most important clusters as the reliably performing SVM with linear kernel. Therefore, it was scrutinised whether the top 5 cluster elements were meaningful regarding the research question (see Appendix N). Eventually, these cluster elements were considered not meaningful to the research topic, hence this filtering was abandoned. Accordingly, the best filtering setup has proven to be the one obtained by the third filtering step, which will be referred to as the 'optimal filtering setup'.

Table 5. The performance of the succeeding models on the third data set with unigrams as features.

	3/I. Correlated features removed		**3/II. Correlated features kept**			**3/III. Correlated features kept with tailored stopword list**		
	SVM linear	SVM radial	Random Forest	LVQ	SVM polynomial	Logistic Regression	Random Forest	SVM polynomial
Balanced accuracy	0.6865	0.6488	0.6310	0.7798	0.7242	0.6310	0.7044	0.7242
F1	0.8814	0.8852	0.8667	0.8929	0.8772	0.8667	0.9000	0.8772

Discussion

The purpose of this study was to examine whether identifying a minority Twitter subpopulation was possible. To this end, social network and tweets (text data) were compared as feature sets. With an F1-score of 0.97 for the best classifier with the social network features, and with 0.90 for that of the text features, the purpose of the study was fulfilled.

Discussion of first data set

The analysis of the first data set aimed to address the research question (**Q1**) of whether the population of Dutch veg*an Twitter users was possible to identify. The data set contained the data of the followers of 24 arbitrarily chosen accounts. Despite being the simplest data collection method, it resulted in assessable predictive performance with most of the applied models. Regarding the predictive performance and comparing it to the above-referred literature, it is declared, that even in case of the simplest network analysis, the F1-score values, as the main measures for predictive performance, of the current research (0.81 – 0.91) exceeded the competing systems (0.60 – 0.86). Moreover, the current research produced more stable results across the classifiers, than the referred ones in all performance measures (F1-score, precision and recall; see Table 6). To conclude, the identification of the Dutch veg*an Twitter subpopulation is possible (**Q1**).

Discussion of the second data set

With the second data set, which contained the social network data of the labelled users, the aim was to test, whether the target population can be predicted based on their social network (**Q2)**. The results show that it is not only possible, but even an increased predictive performance was achieved (F1-score of 0.90 – 0.97) in comparison with the first data set. This also justifies hypothesis 1 (**H1)**, according to which the target group can be predicted based on their social network. Moreover, the predictive performance based on social network (second data set) outperforms both that of the first data set and those of the referred researches. That is, higher and more stable F1-score values were attained (0.90 – 0.97) in the current research than in the competing systems (0.60 – 0.86; see Table 6).

Discussion of the third data set

The analysis of third data set aimed to test whether it is possible to predicting the target population based on their tweets (text data; **Q2**). It was also assessed, whether the predictive performance based on the text features was better than that on the social network (**Q2** and **H3**).Gathering the last 200 tweets of the chosen users and finding the optimal way of data filtering was a time-consuming activity in comparison with the tasks with the first two data sets (hypothesis H4). This is in accordance with Chamberlain, Humby, and Deisenroth (2017), who emphasise that analysing the social network of users is more time- and cost-efficient than analysing their tweets.

The analysis of the tweets required several steps to filter the data in multiple ways. The intermediate analysis proved that the data filtering plays a crucial role in the prediction power. Nevertheless, finding the right combination of possible filters and running all of the analyses several times took a lot of time and effort. In spite of the extra effort, modelling based on the text data resulted in a lower predictive performance than modelling based on the social network.

Based on the chosen performance measures (balanced accuracy and F1-score), keeping the correlating features in the data for the analysis generally increased the predictive performance (balanced accuracy = 0.78, F1-score = 0.89) in comparison with removing them (balanced accuracy = 0.69, F1-score = 0.88). Tailoring the stopword list to the research resulted in even higher F1-score vales (0.90; balanced accuracy = 0.72). When removing the # and @ symbols from the beginning for the words, however, the models either failed to produce assessable results, or when they did, they were not reliable. Based on this finding, the # and @ symbols were retained as it was proven, that they are a special integral property of tweet texts.

Testing whether the combination of words (bigrams) could be a valuable feature to distinguish between the veg*an and non-veg*an people resulted in not assessable predictive performance. Therefore, the data set, in which the # and @ symbols were retained and contained only single words (data set 3/III) was selected as the optimal filtering method.

In sum, predicting the target population is possible based on the users' tweet (text data) (**H2**). Moreover, the current research resulted in higher and more stable performance values (F1-score: 0.87 – 0.90) than those of the referred researches (0.65 – 0.86). However, hypothesis 3 (**H3)** has to be rejected**,** as the predictive performance was not better at the text feature set than that at the social network. This finding contradicts the finding of Chamberlain, Humby, and Deisenroth (2017), although the difference is not substantial.

General discussion

As discussed in the introduction section, the field of social media data analysis is a developing domain without officially agreed rules and/or principles of model use and without methodological standards. Regarding the model selection, despite applying several kinds of classifiers, the predictive performance of the Support Vector Machines appeared to be the best (F1-score range over the data sets: 0.82 – 0.97). Their general popularity for supervised learning tasks seems to be justified for this field of data analysis as well. Besides the Support Vector Machines, the Random Forest classifier performed highly in several instances (F1-

score range over the data sets: 0.87 – 0.90) . Contrariwise, despite being a method specially suitable for the structure of the data, the Fuzzy-Rule Based Classifier with Chi algorithm failed to perform reliably on all data sets.

In sum, the conclusion can be made that it is possible to retrieve the Dutch veg*an Twitter users based on their social network and tweets. As for the results of the current study, they are generally more stable in predictive performance and lead to more accurate results than those of the referred ones (see Table 6). Moreover, regarding the F1-score values, the current research outperformed the reference values network and tweet analysis wise (see Table 6). However, if comparing the predictive performance among the data sets within this research, the best predictive performance was achieved by the extended network analysis (F1-score: 0.90 – 0.97), whilst the text (tweet) analysis resulted in the lowest predictive performance (F1-score: 0.87 – 0.90).

To further elaborate on the findings regarding the lower predictive performance of text data than that of social network data, the theoretical background needs to be recalled about social desirability. For the better understanding, the analogy of *coming out* is proposed, where tweeting about something would refer to the general use of *coming out*, which is a direct and open expression of personal interest or opinion. Following an account would be referred to as *not coming out*, since it is a much less direct, almost confidential expression of interest. Without getting into details of its social psychological background, from the social desirability point of view, it is plausible that finding any minority population is more effectively achievable by a social network analysis than with the help of the tweets (text data). The reason behind it is that emphasising (tweeting about) of belonging to any kind of minority group is generally not socially desirable. Additionally, a great advantage of analysing the social network data is that that type of data is much less protected than the tweets of the users, hence more data is available openly.

Table 6. Predictive performance of the data sets of the current research and the reference studies

	Precision	**Recall**	**(micro) F1-score**
Reference score for network analysis	0.39 – 0.96	0.50 – 0.95	0.60 – 0.86
First data set (base network analysis)	0.81 – 0.91	0.75 – 0.95	0.81 – 0.91
Second data set (extended network analysis)	0.87 – 0.90	0.91 – 0.97	0.90 – 0.97
Reference score for text analysis	0.67 – 0.93	0.45 – 0.98	0.65 – 0.86
Third data set (text analysis) – optimal filtering	0.81 – 0.86	0.89 – 0.96	0.87 – 0.90

Research constraints

The first limitation of the current research is that since the data was collected from the internet, it excluded the possibility of direct – real-life – observation of the sample. Hence, the assumption was made that the online behaviour is a proxy for the offline behaviour. Deducting from the aforementioned assumption, it can also be stated, that the measured values are reflecting attitudes. Hence, when referring to veg*an and non-veg*an population in the study, actually it is about veg*an and non-veg*an minded people who are assumed to be veg*an and non-veg*an in the real-life. Unfortunately, the validity of this assumption cannot be tested by means of social data analysis.

The second limitation is related to the sample. Although it was important that the results could be generalised, collecting online – Twitter – data excludes the non-Twitter users from the sample, thereby causing sampling bias. Admitting that it sounds to be a very strict limitation, given the research topic and the population of interest, the constraint of Twitter users is less serious. The reason is, that being veg*an is a relatively new phenomenon, which means that it is expected to be more popular among the youth compared with the elderly. Since approximately the 37% of the Twitter users is between the age of 18 ad 29, and another 25% is between 30-49[9], the expected sampling bias is low.

The third limitation is the sampling framework. Namely, that the initial data set was collected from veg*an-related accounts. Hence, the probability of sampling meat eaters was much lower than it would have been in a representative sample, ergo the veg*ans were overrepresented in the data sets. However, since the aim of the study was to find the veg*an Twitter users, the interest lay rather in the properties of the veg*an accounts than other ones.

The fourth limitation is the uneven distribution of the positive (veg*an) and negative (non-veg*an) cases in all three samples. Despite the hard definition of the imbalanced data says that whenever the two class numbers are not equal the data is imbalanced, in practice, especially in the field of big data, a data set is called imbalanced when the proportion of positive and negative cases is severely skewed. The distribution of the classes is measured by the imbalance ratio (IR), where a 98:2 IR is deemed to be "fairly common in various real-world scenarios" (Fernandez, del Rio, Chawla & Herrera, 2017, p. 106). In the current research, the IR was roughly 2:1, therefore, no special handling was required. Yet, benchmarks were used for the measurement of the predictive performance, which accounted

[9] https://www.omnicoreagency.com/twitter-statistics/

for the slight imbalance in the data sets. Hence, this limitation did not in fact influence the study.

The fifth limitation is the sample size. Although other authors (Chamberlain, Humby, and Deisenroth, 2017) worked with an equally small proportion of labelled data set[10], it has to be acknowledged, that when the number of labelled cases is low, then there is an increased risk of overfitting. That is, learning the classifiers to recognise patterns at a personal level as they were general trends at the group level. Unfortunately, assessing the severity of this limitation is only possible when comparing the algorithms with other ones that learned on a larger number of labelled cases.

The sixth limitation is that the tweets of the users were downloaded two months after the manual labelling. Therefore, it is possible, that some users changed their diet during those two months, which might have caused bias regarding the labels, and therefore, was the predictive performance lower in case of the text data. However, this scenario seems to be very unlikely.

The seventh limitation is that since the social network models all depend on the accounts that one follows, these models have to be re-evaluated over time. Considering the dynamically changing character of social media, it is highly conceivable, that today's important predictors will be taken over by 'trending' accounts tomorrow. Therefore, in order to ensure the long-time reliability of the models, it is crucial to update them from time to time.

Conclusion

This research intended to filter a hard-to-reach, minority population among social media users in a methodologically correct and generalisable way. To this end, data retrieval and predictive performance were tested along three different approaches. For modelling, supervised machine learning techniques were applied. The class labels, for obtaining high accuracy, were manually assigned to the users according to clear rules. The first approach for data retrieval was to study the followers of arbitrarily chosen Twitter accounts, which were assumed to be followed by the target population. This first approach was used to get the labelled users as well. The second approach was to analyse the social network ('friends') of labelled users. Finally, the third approach was to analyse the text data (tweets) of the labelled users, thereby recognising the different pattern between the target group and others.

[10] 700 million users, 133.000 labelled cases; proportion = 0.0002. In the current study: 67.665 users, 165 labelled cases; proportion = 0.0024

The research resulted in assessable predictive performance in all of the three approaches. Therefore, the conclusion can be drawn that retrieving the population of Dutch veg*an Twitter users is possible by both social network and text analysis. Moreover, high predictive performance is achievable based on all of the three approaches, whilst the three approaches are approximately equally efficient. However, to reach an optimal outcome, a strict adherence to high methodological standards was required.

Further research

Classifying the Dutch population

To complete this research, the next step would be the classification of the Dutch Twitter population, that is, identifying all Dutch, veg*an Twitter users. Unfortunately, this step can only be taken theoretically. The reason is that the complete list of Dutch Twitter accounts has not been retrieved yet, and even if it had been, analysing the social network of all Dutch Twitter users would be out of the time range of this thesis. If there were sufficient resources available to carry on with the research, the next steps would look as follows:

Considering that the network analysis proved to be more reliable than the text analysis for the prediction of the Dutch veg*ans, and taking the cost-benefit ratio into account, the network analysis would be the best method to apply. Although using an ensemble method might result in even increased predictive performance, for the sake of efficiency, social network analysis would be preferred. The initial step would be to collect the followers of the two Dutch veg*an associations (*Nlvegan* and *vegetaiersbond*) as these appeared to be reliable predictors for the target population. Once the followers have been collected, the users who follow both accounts would be classified as veg*ans with high confidence. Thereafter, their friends network could be analysed and the most important features could be extracted. Based on these features, a given number of additional users could be classified. Once these users have been classified, their friends network could be analysed and the most important features could be re-evaluated. When having a list of the newly evaluated features, again, a given number of additional users could be classified. By such an iterative way of classification, all Dutch Twitter accounts could be labelled as veg*an or not. As soon as the target population has been identified, they could be studied extensively. For example, the amount of veg*an people across the Netherlands could be estimated, and based on the geotags, the dispersion of the population could be measured. Additionally, their demographic properties and motivating factors could be scrutinised.

General recommendations for further social media researches

The intention of the study was to present a replicable and methodologically correct example for analysis of social media data. The key aspects of the research: First, knowing the offline behaviour of the target population is an important initial step, which should not be bypassed. Second, accounting for the possible biases beforehand (see listing in Ruths & Pfeffer, 2014) and taking appropriate steps to prevent them or correct them afterwards is crucial for the reliability of the result. Third, the selected models should be always suitable for the data properties. Fourth, in case of tweet (text) analysis, special attention should be devoted to the text filtering phase as it has been proven to effect the predictive performance seriously.

General application

As the above-introduced techniques are generalisable, and can be applied for identifying any kind of minority population, let it be the group of cancer patients, homosexuals, victims of abuse, and so forth, only the fantasy can limit the possible adaptations. Hopefully, the provided framework will help to understand and in case of need, support these groups of our society.

References

Aiello, S., Eckstrand, E., Fu, A., Landry, M., & Aboyoun, P. (2017). *Machine learning with R and H2O.* Available from: http://h2o2016.wpengine.com/wpcontent/themes/h2o2016/images/resources/ Rbooklet.pdf

Benoit, K. (2017). *Quanteda: Quantitative Analysis of Textual Data*. R package version 0.99.22. Available from: https://CRAN.R-project.org/package=quanteda

Beel, J., Gipp, B., Langer, S. & Breitinger, C. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries, 17*(4)*, 305–338.*

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Infromation Engineering and Applications, 3*(10), 27-39. Retrieved from: https://eva.fing.edu.uy/pluginfile.php/69453/mod_resource/content/1/ 7633-10048-1-PB.pdf

Buller, D. B. & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203-242. Doi: 10.1111/j.1468-2885.1996.tb00132.x

Caspi, A., & Gorsky, P. (2006). Online deception: prevalence, motivation, and emotion. *Cyberpsychology & Behavior, 9*(1)*,* 54-59. Doi: 10.1089/cpb.2006.9.54

Cesare, N., Grant, C., & Nsoesi, E. O. (2017). *Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices.* Retrieved from: https://arxiv.org/ftp/arxiv/papers/1702/1702.01807.pdf

Chamberlain, B. P., Humby, C., & Deisenroth, M. P. (2017). Probabilistic Inference of Twitter Users' Age based on What They Follow. *Proceedings of the European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases.* Retrieved from: https://arxiv.org/pdf/1601.04621.pdf

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent od psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354. Doi: 10.1037/h0047358

Culotta, A., Ravi, N. K., Cutler, J. (2016). Predicting Twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research, 55,* 389-408. Doi: 10.1613/jair.4935

Elkano, M., Galar, M., Sanz, J., & Bustince, H. (in press). CHI-BD: A fuzzy rule-based classification system for big data classification problems. *Fuzzy Sets and Systems*. Doi: 10.1016/j.fss.2017.07.003

Fernandez, A., del Rio, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems, 3*, 105-120. Doi: 10.1007/s40747-017-0037-9

Fox, N., & Ward, K. (2008). Health, ethics and environment: a 31eft31iour31e study of vegetarian motivations. *Apetite, 50,* 422-429.

Gentry, J. (2015). 31eft31io: R based Twitter client. R package version 1.1.9. Available from: https://CRAN.R-project.org/package=31eft31io

Golder, S. A., & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research*. Annual Review of Sociology, 40*(1), 129-152.

Grossman, M. (2017). Study of social media users: the relation between online deception, Machiavellian personality, self-esteem, and social desirability (Doctoral Dissertation). Retrieved from ProQuest. (ProQuest number: 10617442).

Hoffman, S. R., Stallings, S. F., Bessinger, R. C., & Brooks, G. T. (2013). Differences between health and ethical vegetarians. Strength of conviction, nutrition knowledge, dietary restriction, and duration of adherence. *Appetite, 65*, 139-144.

Kaden, M, Riedel, M., Hermann, W., & Villmann, T. (2015). Border-sensitive learning in generalized Learning Vector Quantization: an alternative to Support Vector Machines. *Soft Computing, 19*(9), 2423-2434. Doi: 10.1007/s00500-014-1496-1

Kashima, Y., McKintyre, A., & Clifford, P. (1998) The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology, 1,* 289-313.

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., Natsev, A., & Xie, L. (2012). Social media use by government: From the routine to the critical. *Government Infromation Quarterly, 29,* 480-491.

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Jupyter Development Team (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. Available from: http://ebooks.iospress.nl/publication/42900, doi: *10.3233/978-1-61499-649-1-87*

Kowsari, K., Bari, N., Vichr, R., & Goodarzi, F. A. (2018). FSL-BM: Fuzzy supervised learning with binary meta-feature for classification. Proceedings from FICC ’18: *Future of Information and Communications Conference.* Singapore.

Kuhn, M. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). *Caret: Classification and Regression Training*. R package version 6.0-77. Available from: https://CRAN.R-project.org/package=caret

Martinez, O., Wu, E., Shultz, A. Z., Capote, J., Rios, J. L., Sandfort, T., Manusov, J., Ovejero, H., Caballo-Dieguez, A., Baraya, S. C., Moya, E., Matos, J. L., DelaCruz, J. J., Remien, R. H., & Rhodes, S. D. (2014). Still a Hard-to-Reach Population? Using social media to recruit latino gay couples for an HIV intervention adaptation study. *Journal of Medical Internet Research, 6*(4). Doi: : 10.2196/jmir.3311

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017*). E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.6-8. Available from: https://CRAN.R-project.org/package=e1071

Mullee, A., Vermeire, L., Vanaelst, B., Mullie, P., Deriemaeker, P., Leenaert, T., De Henauw, S., Dunne, Aoibheann, Gunter, M. J., Clarys, P., & Huybrechts, I. (2017). Vegetarianism and meat consumption: A comparison of attitudes and beliefs between vegetarian, semi-vegetarian, and omnivorous subjects in Belgium. *Appetite*, 114, 299-305. Doi: 10.1016/j.appet.2017.03.052

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). “How Old do You Think I am?” A Study of Language and Age in Twitter. Proceedings from ICWSM ’13: *The 7th International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Web and Social Media.* Boston: USA.

Nova, D., & Estevez, P. A. (2014). A review of Learning Vector Quantization classifiers. *Neural Computing and Applications, 25*(3-4), 511-524. Doi: 10.1007/s00521-013-1535-3

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of I Bayes text classifiers. Proceedings from ICML ’03: *The 20th International Conference on Machine Learning.* Washington, DC: USA.

Roesslein, J. (2009). Tweepy. Python package version 3.5.0. Available from: www.tweepy.org

Rstudio Team (2016). Rstudio: Integrated Development for R. Rstudio, Inc., Boston, MA: USA. Available from: http://www.rstudio.com/

Ruths, D., & Pfeffer, J. (2014). Social medi for large studies of 34eft34iour – Large-scale studoes of human 34eft34iour in social media need to be held to higher methodological standards. *Science, 346*(6213), 1063-1064. Doi:10.1126/science.346.6213.1063

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., Ungar, L. H., & Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. Proceedings from EMNLP '14: *Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar.

Sloan, L., & Quan-Haase, A. (2017). T*he SAGE Handbook of Social Media Research Methods.* Los Angeles, California: SAGE Publications Inc.

Tabachnick, B. G. & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Volkova, S., van Durme, B., Yarowsky, D., & Bachrach, Y. (2015). Social media predictive analytics. Proceedings from NAACL-HLT '15: *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*. Denver, CO: USA.

Williams, M. L., Burnap, P., Sloan, L. (2017). Crime sensing with big data: the affordances and limitations of using open-source communications to estimate crime patterns. *British Journal Of Criminology, 57*(2), 320-340. Doi:10.1093/bjc/azw031

Zhang, H. (2004). The optimality of I Bayes. Proceedings from FLAIRS '04: *The 17th International Florida Artificial Intelligence Research Society Conference.* South Beach, FL: USA.

Appendix A

List of the chosen Twitter accounts

NVV Veganisme NL – *@Nlvegan* – Dutch Association for Veganism

PETA Nederland – *@PETANederand* – Animal rights association

Viva Las Vega’s – *@vivalasvega* – Association for a vegan lifestyle

Vegetariersbond – *@vegetariersbond* – Association for vegetarianism

PvdD – *@PartijvdDieren* – political party for the animals

Vegetarische Slager – *@VegaSlager* – veg(etari)an food brand

Vivera – *@vivavivera* – veg(etari)an food brand

Jacinta Bokma – *@devegetarier* – vegan celebrity food blog

Lisette Kreischer – *@VeggieinPumps* – vegan celebrity food blog

Dyana Loehr – *@DyanaLoehr* – vegan celebrity food blog

Lisa Steltenpool – *@LisaSteltenpool* – vegan celebrity food blog

Plantaardigheidjes – *@World_of_Dani* – vegan food blog

plantaardigLeven.nl – *@plantaardigETEN* – vegan food blog

Veg-O-Holic – *@Vegoholic* – vegetarian food blog

Vegafit – *@VegafitNL* – vegan food blog

Vegatopia – *@Vegatopia* – vegan food blog

VeganChallenge – *@30dagenvegan* – foodblog of the 30 days vegan challange

Veganistisch Koken – *@VeganKock* – vegan food blog

Lekker vegetarisch – *@HeavenofDelight* – vegan news blog

Vegetafel – *@Vegetafel* – vegan event organiser

VegFestNL – *@VegFestNL* – vegan event blog

Veggie Fair – *@VeggieFair*– veg(etari)an event

Veggies On Fire – *@VeggiesOnFire* – plant based restaurant

Veggie 4U – *@Veggie4you* – vegan shop

Appendix B

The collected metadata of the Twitter users

user.id – the Twitter ID number

user.screen_name – the screen name

user.name – the user name

user.location – the location of the user

user.profile_location – the location of the profile

user.time_zone – the time zone of the user

user.description – the profile description of the user

user.status.text – the actual state text of the user at the time of the data downloading

user.lang – the language setup, on which the Twitter is displayed

user.followers_count – the number of the followers of the given user

user.created_at – the date and time when the account was created

Appendix C

The performance of the models on the first data set

	Logistic Regression	**Fuzzy classifier**	**Boosted Tree**	**Random Forest**	**LVQ**	**SVM radial**	**SVM linear**	**SVM polynomial**
Top 5 features	vegetariersbond, PartijvdDieren, 30dagenvegan vivalasvega, vivavivera	Nlvegan, vegetariersbond, vivalasvega, PartijvdDieren, HeavenofDelight	Nlvegan, vegetariersbond, vivalasvega, PartijvdDieren, HeavenofDelight	Nlvegan, PartijvdDieren, vegetariersbond vivavivera, VegaSlager	Nlvegan, vegetariersbond, vivalasvega PartijvdDieren, HeavenofDelight	Nlvegan, vegetariersbond, vivalasvega PartijvdDieren, HeavenofDelight	Nlvegan, vegetariersbond, vivalasvega PartijvdDieren, HeavenofDelight	Nlvegan, vegetariersbond, vivalasvega, PartijvdDieren, HeavenofDelight
Balanced accuracy	0.7962	0.5385	0.8087	0.7452	0.6308	0.6413	0.7596	0.6288
McNemar's test	1.000	0.0015	1.000	1.000	1.000	0.0269	0.09609	0.0704
Recall	0.9000	1.000	0.9250	0.8750	0.8000	0.9750	0.7500	0.9500
Precision	0.9000	0.7692	0.9024	0.8750	0.8205	0.8125	0.9091	0.8085
F1-score	0.9000	0.8696	0.9136	0.8750	0.8101	0.8864	0.8219	0.8736

Appendix D

List of user ID numbers, which were followed at least by the 25% of the veg*an users

130801203, 15207550, 20710359, 118675795, 44962002, 242337797, 252604300, 50600050, 46249920, 45600930, 139400870, 303744759, 398238378, 223471353, 7174972, 23577041, 2493701, 85434447, 33944106, 972167018, 73620907

Appendix E

List of accounts, which were followed at least by the 25% of the non-veg*an users

130801203, 148287455, 226931469, 102950029, 136193493, 334003382, 16891107, 757210334, 127342372, 176436528, 291702948, 34627041, 38407379, 45600930, 54151280, 102402934, 209977998, 214466767, 259691725, 69233745, 81573850, 88302046, 138752515, 140483194, 15207550, 174958019, 342144061, 45213585, 476683222, 53628337, 580875349, 73970612, 87017921, 88940763, 92561432, 105735168, 125127787, 134784499, 210349500, 22477925, 292299338, 41215408, 106769377, 1103335105, 126913615, 130944392, 145981361, 146068604, 160965168, 174950107, 209608901, 223471353, 266087692, 274948633, 300204490, 309064220, 30949745, 356129820, 42849490, 434896581, 50280503, 66840275, 96986572, 1017776419, 103423196, 114463320, 12654112, 142640094, 145993108, 150282681, 162831275, 19182978, 297583376, 38518252, 391567080, 413972878, 455013786, 48424104

Appendix F

User accounts used as features in the second data set

VegaSlager, mariannethieme, PartijvdDieren, Nlvegan, Vegatopia, vegetariersbond, vivalasvega, HeavenofDelight, devegetarier, WakkerDier, estherouwehand, AnimalsToday_nl, merelwildschut, vivavivera, NOS, wnfnederland, Nunl, VeggieinPumps, bontvoordieren, 30dagenvegan, vegalifeNL, boerenfluitjes, okvleesnl, FoodReporter, FoodiesMagazine, CulyNL, foodlog_nl, MeatYourOwn, delibrije, MeatCo1, ELLEeten, foodinspiration, foodwatch_nl, EtenMetMara, demoslager, MacvanDinther, worstmakers, samuellevie, Kokenmetkarin, LekkerTafelen, AstridsTaste, FabulousFoodFan, Spinazieacademi, NuijtenRob, vossius, Streekbox, Desemenzo, TopenVers, KvW, 24Kitchen, deliciousnl, onnokleyn, ronblaauw, wateetjanneke, pfkalfsvlees, chickslovefood, Talkinfood, KNSvoorslagers, dickfoodlognl, vandalenvlees, Spoelder1885, Worstmaker, ZTRDG, PetravanHaandel, yvettevanboven, EetWeters, CasaForesta, renepluijm, paulineskeuken, horeca_gids, Vleesnl, NieuwVers, WillemenDrees, SVO_Opleidingen, Culinette, Hermandenblijk, Puuruiteten, FelixWilbrink, eetschrijver, slagerspassie, wvanlaarhoven, BionextTweets, Wateetons, smaakvrienden, Hanssteenbergen, WeekvandeSmaak, HeijdraVleesvee, dwdd, LimousinVlees, Dierbescherming, KoeOpAvontuur, Hela_slagerij, SFYN_NL, JorisLohman

Appendix G

The performance of the models on the second data set

	Logistic Regression	Fuzzy classifier	Boosted Tree	Random Forest	LVQ	NB	SVM linear	SVM polynomial
Top 5 features	CulyNL, KNSvoorslagers, Worstmaker, renepluijm, KvW	CulyNL, FoodReporter, FoodiesMagazine, Nlvegan, boerenfluitjes	CulyNL, FoodReporter, FoodiesMagazine, Nlvegan, boerenfluitjes	KNSvoorslagers, LekkerTafelen, CulyNL, Hela_slagerij, FoodiesMagazine	CulyNL, FoodiesMagazine FoodReporter, Nlvegan, boerenfluitjes	CulyNL, FoodiesMagazin FoodReporter, Nlvegan, boerenfluitjes	CulyNL, FoodiesMagazin FoodReporter, Nlvegan, boerenfluitjes	CulyNL, FoodiesMagazin FoodReporter, Nlvegan, boerenfluitjes
Balanced accuracy	0.8190	0.5417	0.8333	0.7905	0.7774	0.6667	0.8333	0.8190
McNemar’s test	0.3711	0.002569	0.1336	1.000	0.2207	0.0133	0.1336	0.3711
Recall	0.9714	1.000	1.000	0.9143	0.9714	1.000	1.000	0.9714
Precision	0.8947	0.7609	0.8974	0.8889	0.8718	0.8140	0.8974	0.8947
F1-score	0.9315	0.8642	0.9459	0.9014	0.9189	0.8974	0.9667	0.9315

Appendix H

List of stopwords for the third data set

I, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, would, should, could, ought, I'm, you're, he's, she's, it's, we're, they're, I've, you've, we've, they've, I'd, you'd, he'd, she'd, we'd, they'd, I'll, you'll, he'll, she'll, we'll, they'll, isn't, aren't, wasn't, weren't, hasn't, haven't, hadn't, doesn't, don't, didn't, won't, wouldn't, shan't, shouldn't, can't, cannot, couldn't, mustn't, let's, that's, who's, what's, here's, there's, when's, where's, why's, how's, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, de, en, van, ik, te, dat, die, in, een, hij, het, niet, zijn, is, was, op, aan, met, als, voor, had, er, maar, om, hem, dan, zou, of, wat, mijn, men, dit, zo, door, over, ze, zich, bij, ook, tot, je, mij, uit, der, daar, haar, naar, heb, hoe, heeft, hebben, deze, u, want, nog, zal, me, zij, nu, ge, geen, omdat, iets, worden, toch, al, waren, veel, meer, doen, toen, moet, ben, zonder, kan, hun, dus, alles, onder, ja, eens, hier, wie, werd, altijd, doch, wordt, wezen, kunnen, ons, zelf, tegen, na, reeds, wil, kon, niets, uw, iemand, geweest, andere, via, http, co

Appendix I

The performance of the models on data set 3/I. – Without correlated features

	Logistic Regression	**Fuzzy classifier**	**Boosted Tree**	**Random Forest**	**NB**	**LVQ**	**SVM radial**	**SVM linear**	**SVM polynomial**
Top 5 features							1, weer, nieuwe, s, vegan	1, weer, nieuwe, s, vegan	1, weer, nieuwe, s, vegan
Balanced accuracy							0.5556	0.6865	0.6488
McNemar's test							0.0133	0.4497	0.1306
Recall							1.0000	0.9286	0.9643
Precision							0.7778	0.8387	0.8182
F1-score	Broke down	Broke down	Broke down	Predicted only veg*ans	Predicted only veg*ans	Predicted only veg*ans	0.8750	0.8814	0.8852

Appendix J

The performance of the models on data set 3/II. – With correlated features

	Logistic Regression	Fuzzy classifier	Boosted Tree	Random Forest	NB	LVQ	SVM radial	SVM linear	SVM polynomial
Top 5 features				Onze, wij, lekker, 2016, rode		Onze, wij, week, goed, vlees	Onze, wij, week, goed, vlees	Onze, wij, week, goed vlees	Onze, wij, week, goed, vlees
Balanced accuracy				0.6310		0.7798	0.5556	0.6111	0.7242
McNemar's test				0.2888		1.000	0.0133	0.0233	1.000
Recall				0.9268		0.8929	1.000	1.000	0.8929
Precision				0.8125		0.8929	0.7778	0.7897	0.8621
F1-score	Broke down	Broke down	Broke down	0.8667	Predicted only veg*ans	0.8929	0.8750	0.8889	0.8772

Appendix K

The performance of the models on the third data set 3/III. – With tailored stopword list

	Logistic Regression	**Fuzzy classifier**	**Boosted Tree**	**Random Forest**	**NB**	**LVQ**	**SVM radial**	**SVM linear**	**SVM polynomial**
Top 5 features	Kip, sylviawitteman, bezorgen, foto, 17			Lekker, vlees, 2016, weer, 1, mooie				Goed, wel, niet, vlees, weer	Goed, wel, niet, vlees, weer
Balanced accuracy	0.6310			0.7044				0.6111	0.7242
McNemar's test	0.2888			0.2207				0.0233	1.0000
Recall	0.9286			0.9643				1.000	0.8929
Precision	0.8125			0.8438				0.8000	0.8621
F1	0.8667	Broke down	Broke down	0.9000	Predicted only veg*ans	Predicted only veg*ans	Predicted only veg*ans	0.8889	0.8772

Appendix L

The performance of the models on data set 3/IV. – Without # and @ characters

	Logistic Regression	**Fuzzy classifier**	**Boosted Tree**	**Random Forest**	**NB**	**LVQ**	**SVM radial**	**SVM linear**	**SVM polynomial**
Top 5 features				Vlees, week, lekkerite, rode, texel			Week, vlees, niet, vandaag, vegan	Week, vlees, niet, vandaag, vegan	Week, vlees, niet, vandaag, vegan
Balanced accuracy				0.6500			0.5500	0.5333	0.8000
McNemar's test				0.02334			0.0077	0.0269	0.0269
Recall				1.000			1.000	0.9667	0.7000
Precision				0.8108			0.7692	0.7632	0.9545
F1	Predicted only veg*ans	Broke down	Did not converge	0.8955	Predicted only veg*ans	Predicted only veg*ans	0.8696	0.8529	0.8077

Appendix M

The performance of the models on data set 3/V. – Bigrams

	Logistic Regression	Fuzzy classifier	Boosted Tree	Random Forest	NB	LVQ	SVM radial	SVM linear	SVM polynomial
Top 5 features	rode kool, wel leuk, twee weken, wensen iedereen, per kilo			rode kool, wel leuk, harte welkom, volgende week, nieuwe website	volgende week, facebook geplast, harte welkom, vandaag weer, rode kool	volgende week, facebook geplast, harte welkom, vandaag weer, per stuk	volgende week, facebook geplast, harte welkom, vandaag weer, per stuk	volgende week, facebook geplast, harte welkom, vandaag weer, per stuk	volgende week, facebook geplast, harte welkom, vandaag weer, per stuk
Balanced accuracy	0.6103			0.5270					
McNemar's test	0.0269			0.0094					
Recall	0.9706			0.9706					
Precision	0.7857			0.7500					
F1	0.8684	Breaks down	Does not converge	0.8462	Predicts only veg*ans	Predicts only veg*ans	Predicts only veg*ans	Predicts only veg*ans	Predicts only veg*ans

Appendix N

The performance of the models on data set 3/VI. - Word clusters, the elements of the top 5 clusters are detailed below

	Logistic Regression	**Fuzzy classifier**	**Boosted Tree**	**Random Forest**	**NB**	**LVQ**	**SVM radial**	**SVM linear**	**SVM polynomial**
Top 5 clusters			100, 31, 21, 25, 24					100, 31, 21, 25, 24,	100, 31, 21, 25, 24,
Balanced accuracy			0.5933					0.7421	0.5556
McNemar's test			0.0771					0.6831	0.0133
Recall			0.9643					0.9286	1.000
Precision			0.7941					0.8667	0.7778
F1	Predicted only veg*ans	Broke down	0.8710	Predicted only veg*ans	Predicted only veg*ans	Predicted only veg*ans	Predicted only veg*ans	0.8966	0.8750

Cluster	**Elements**
100	#woerden, #harmelen, #ouderen, @zgsintmaarten
31	klant, @jpoetijn, boeken, mannen, #carredebat
21	12, fiets, @xlottexx, sluit, onderweg, par, trein, @ovchipkaart, #druk, beginnen, manege
25	#beleggen, #investing, #rabobank, #esg, @aldertv, #em, #robotics, #fondsen, #mobius, actueel, index
24	#moscow, baas, russen