The Detection of Fake Messages using Machine Learning

Maarten S. Looijenga
University of Twente
PO Box 217, 7500 AE Enschede
The Netherlands
m.s.looijenga@student.utwente.nl

ABSTRACT

This research investigates how fake messages are used on Twitter during the Dutch election of 2012. It researches the performance of 8 supervised Machine Learning classifiers on a Twitter dataset. We provide that the Decision Tree algorithm perform best on the used dataset, with an F-Score of 88%. In total, 613.033 tweets were classified, of which 328.897 were classified as true, and 284.136 tweets were classified as false. Through a qualitative content analysis of false tweets sent during the election, distinctive features and characteristics of false content have been found and grouped into six different categories.

Keywords

Machine Learning, politics, social media, automated content analysis, fake news

1. INTRODUCTION

Many people use social media as a communication tool. In the last few years, social media has grown extensively. Our research focusses on the social media platform Twitter. Twitter is a social media networking site. In The Netherlands alone, Twitter has approximately 2.8 million users, of whom 1.0 million people use Twitter on a daily basis [25]. People communicate with each other through tweets, short text messages with a maximum of 280 characters. Social media can be used as a marketing tool to reach many people quickly. People do not only use the medium to share events of their lives, but also to share their opinions about many topics. Messages on Twitter can be read by almost everyone who wants to read it. Tweets can be read by nearly everyone who has the urge to read those messages. [24]. Content can be relayed among users with no significant third-party filtering, fact-checking, or editorial judgment. An individual user with no track record or reputation can in some cases reach as many readers as Fox News, CNN, or the New York Times [1].

In the last years, privacy concerns about social media have risen. At the beginning of 2018, the British news channel Channel 4 published an article about the influence of data-analytics company Cambridge Analytica on the USA presidential elections of November 8th, 2016 [26]. Cambridge Analytica has been accused of obtaining data on 50 million Facebook users for marketing purposes [11]. They collected the data via means that deceived both the users and Facebook. The company claimed it could develop psychological profiles of consumers and voters which was a "secret sauce" it used to sway voters more effectively than traditional advertising could [18].

Not only the USA presidential election of 2016 was influenced through extensive data analytics by Cambridge Analytica. Allegations have been made towards the influence of Cambridge Analytica with the United Kingdom European Union membership referendum of 2016 [18][27][28][29]. Chris Wylie, former director of research at Cambridge Analytica and a company whistle-blower, said that "a Canadian business with ties to Cambridge Analytica's parent company, SCL Group, also provided analysis for the Vote Leave campaign ahead of the 2016 Brexit referendum. This research […] likely breached the U.K.'s strict campaign financing laws and may have helped to sway the final Brexit outcome." [25].

The negative campaign messages spread by Cambridge Analytica do not necessarily have to be true. Researchers claim fake news was extensively used to manipulate the outcome of the election [1]. Fake news is defined as "news articles that are intentionally and verifiably false, and could mislead readers."[1]. Many people who see fake news stories report that they believe them [6].

In this research, we will investigate how fake messages can be detected using machine learning. The research will focus on the Dutch election of 2012. A Machine Learning algorithm will be developed to identify untrue content on Twitter. The research will focus on the Dutch population, who used the social media platform Twitter during the Dutch 2012 election. To investigate to what extent fake messages have been used during the Dutch election of 2012, we formulated the following research questions:

- Can we train a classifier to detect potential fake media regarding the Dutch election of 2012?
- What kind of fake messages have been used during the Dutch election period of 2012 on Twitter?

This paper is structured as follows: First, we will describe a literature review, in which our research is compared to already existing research. Second, we will explain the research design and method in detail. Then, the results of both the classification of the Machine Learning classifiers and the qualitative content analysis are described. Finally, the results of the research are discussed, even as the limitations and possible further work.

For this research, we used an existing Twitter dataset. The database consists of tweets posted around the Dutch election of September 12, 2012. The classifier was trained on only the text

29th Twente Student Conference on IT, Jun. 6th, 2018, Enschede, The Netherlands.

of the tweet. The dataset was an existing dataset, gathered using relevant hashtags, like #CDA or #TK2012, in which CDA is the abbreviation of a political party, while TK2012 stands for "Tweede Kamer 2012", referencing to the 2012 parliament election. The data is cleaned to only investigate tweets that are about the Dutch election. An example of a tweet in the dataset:

"Ik heb mijn stem uitgebracht! voor een sterk midden en einde stilstand. Juist nu. #D66 #ikstemd66 http://t.co/b8Y2aYwL".

A sample of 300 tweets was created from the corpus and manually labelled. The data were divided into two different classes: True (1), and False (0). They are defined in section 3.2. Eight different classifiers were trained and compared. The Decision Tree algorithm is performing best on the used dataset, with an F-Score of 88%. This algorithm was used to classify 613.033 tweets, of which 328.897 were classified as true, and 284.136 tweets were classified as false. Through a qualitative content analysis of tweets sent during the election, distinctive features and characteristics of malicious content have been found and grouped into six different categories, which can be found in section 4.2.

2. RELATED WORK

In this chapter, we will describe a literature review, in which our research is compared to already existing research. Research has already been conducted on the detection of nonfactual content on social media, the detection of bots on social media and the influence of persuasive messages on specific elections using social media. First, an overview of the related work will be given. Second, the related work will be compared with our research.

Research has been conducted on the detection of nonfactual content on social media. Keretna, Hossny, & Creighton [17] investigated the possibility of an algorithm that automatically identifies the user identity on Twitter through text mining. It verified the owners of social media accounts, to eliminate the effect of any fake user accounts on people's perception. The algorithm was based on write-print, a writing style biometric. Boididou et al. [2] focussed on the problem of misleading visual content on Twitter. When social media users are posting pictures on Twitter with a description, this description does not have to be true. Boididou et al. [2] discovered that pictures were posted on Twitter with a false description. For example, a photo of a fake shark swimming in a flooded street was used several times after major hurricanes in the USA. They developed a system that supports the automatic classification of multimedia Twitter posts in the categories 'credible' or 'misleading'. The categorisation is based on the text of the message and the user profile that posted the message.

Cresci et al. [7] tried to detect fake Twitter followers efficiently. They tried to identify false users that only were created for the sake of following. These accounts were not used to post (false) messages, but only to follow, like or retweet messages on social media to enhance the popularity of the followed user or topic. Cresci et al. [7] evaluated multiple rulesets that to access it strength in discriminating fake followers. They build a classifier that consists of rules proposed by Academia and Media, containing methodologies for spam and bot detection, to detect anomalous Twitter accounts, in combination with a trained Machine Learning algorithm.

Research has been carried out on the use of social media and the influence of persuasive messages on specific elections [5][10][14][26]. Spierings [26] researched the Dutch elections of 2010 and 2012. The study examined why, when and how political parties had used social media during their campaigning. They investigated if Web 2.0 levels the political playing field or if they mirror existing inequalities between parties. Hosch-Dayican et al. [14] investigated how online citizens persuade fellow voters during the Dutch election of 2012. They analysed the way election campaigns are conducted on Twitter by citizens accounts. During an election campaign, Twitter can be used by voters to convince a fellow voter to vote in favour or against a particular party or leader.

Much research about the automated detection of fake content has already been performed. However, many researchers focus their investigation on the detection of fake users, also known as bots. They try to identify this by looking at the user account that has posted the message. Our classifier is trained on only the textual content of a tweet, ignoring the user account. It would be interesting to built a classifier that can analyse both textual content of the tweet and the user that sent the tweet. Unfortunately, the dataset used contained twitter messages of 2012. These messages are six years old. This meant that much information about the user, such as Twitter followers and number of sent messages, could not be retrieved. Therefore, the Machine Learning algorithm was trained on only the content of the twitter message.

Analysis using Machine Learning algorithms on only text messages has been done before, for instance, Hosch-Dayican et al. [14]. However, they only used one specific Machine Learning algorithm, while we will analyse and compare eight different Machine Learning algorithms and using the best performing algorithm to classify our dataset.

3. RESEARCH DESIGN AND METHOD

In this chapter, we elaborate on the research design and method used in our research. In section 3.1, the data selection and gathering process are explained. In section 3.2, the training process of the classifier is discussed, illustrating the data sampling method, the pre-processing method of the data and the implementation of the different classifiers are discussed. In section 3.3, the process for the qualitative content analysis is presented.

3.1 Data Selection and Gathering

For this research, we used an existing Twitter dataset. The database consists of tweets posted around the election of September 12, 2012. The tweets have been collected by Hosch-Dayican et al. [14]. They researched how online citizens persuaded fellow voters in the Dutch election of 2012. They used the logic of the snowball sampling method to gather relevant hashtags. A hashtag—written with a # symbol—is used to index messages on Twitter. It allows people to follow topics easily according to their interests [15]. The snowball sampling method is based on referrals from initial subjects to generate additional subjects [12]. Primary data sources nominate other data sources to be used in research [9]. The collection started with a list of 19 hashtags of selected parties and their candidates, but also about media events, actual issues and general election hashtags [14]. A script extracted other tags present in mined tweets, to which a relevance was assigned. Once a tag passed a certain threshold, it was added to the list of tags and used to collect new tags.

3.2 Training of Classifier

We decided to perform an automated content analysis on the collected tweets. We used Machine Learning algorithms in this process. Machine Learning is an area of Artificial Intelligence which allows machines to learn from data without guidance. It

gives computers the ability to learn without being explicitly programmed [22].

3.2.1 Data sampling

A classifier has been built to identify potentially fake messages on Twitter. The classifier has been developed using supervised Machine Learning algorithms. Supervised Machine Learning algorithms need training data to teach the model how to behave. We made a training set to train the algorithm. This set contains 300 tweets from the dataset and was manually labelled. Two different classes have been used to label the data. The first class are true messages. These messages are correct messages sent by humans. These messages were labelled with the integer '1'. The other class are false messages. These messages are sent by bots or are incorrect and misleading. These messages were labelled with the integer '0'.

The tweets for the sample have been randomly selected and labelled by the researcher. First, 150 true messages were gathered. These messages were taken from a dataset by Hosch-Dayican et al. [14]. This dataset was used by [14] for their research about the persuasion of fellow voters on Twitter. The dataset was cleaned for their study. An example of a true twitter message:

"Interessante bijeenkomst over decentralisaties in Assen #pvda #lokaalsociaal".

Another 150 tweets from the corpus have been gathered. These messages were considered sent by bots or untrue. These messages were determined to be false, using the Camisani-Calzolari rule set [4], which can be found in Appendix 1. How more rules of this ruleset were considered invalid, the higher the possibility that the tweet was false. To ensure a high validity on the datasample, many tweets have been tested with the Camisani-Calzolari rule set [4]. An example of a false twitter message:

"GELDNOOD? #tk2012 http://t.co/fXe2bB2N http://t.co/Wn5uOr8t"

The 300 collected and labelled messaged have been used as training- and testset for the training of the supervised Machine Learning algorithms.

3.2.2 Pre-processing of data

Data requires special preparation before it can be understood by Machine Learning algorithms. Raw data cannot be fed directly to the algorithm. Most algorithms need integers or floats as input, while tweets are strings. Also, most algorithms expect numerical feature vectors with a fixed size rather than raw texts with a variable length, like tweets [25].

Data needs to be cleaned to increase the reliability of the Machine Learning algorithm [21]. Punctuation was removed to reduce each comment to purely words. Some algorithms treat words followed by punctuation as stand-alone words. For the same reason, special characters were removed. All words were converted to lowercase. The algorithm does not consider two similar words with different capitalisation as equal.

The Scikit-learn package provides utilities for the pre-processing of data. We used The Bag of Words Representation Model. It provides tools for tokenising, counting, normalising and the vectorisation of data. The model throws away all the order information in the messages, but looks at the occurrence of words in a document. Each unique word is assigned with a unique number, a token. Stopwords have not been removed. These type of words will emerge many times but are not meaningful in the encoded vectors [3]. However, since this method looks at the occurrence of words in the document, this is already taken into account.

We will calculate the word frequencies, an alternative to the standard Bag of Words implementation. The function TfidfVectorizer is used [25]. The function is based on 'term frequency-inverse document frequency', shortened TF-IDF. This vectorisation method is mostly used for text analyses with a large text corpus. For each word in the corpus, a floating point value will be calculated. The formula for the calculation is:

$$\text{tf-idf(t,d)} = \text{tf(t,d)} \times \text{idf(t)}$$

where *idf(t)* is:

$$\text{idf}(t) = log\frac{1+n_d}{1+\text{df}(d,t)} + 1.$$

The term frequency, the sum of total occurrences of a word in a given dataset, is multiplied with the *idf* component. N_d is the total number of tweets and *df(d,t)* is the sum of tweets in which the term occurs.

The data is also normalised. Our dataset is normalised using the Euclidean norm:

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+\cdots+{v_n}^2}}$$

3.2.3 Machine Training: The Classifier

For this research, eight different supervised Machine Learning algorithms have been analysed. These Machine Learning algorithms are trained and tested with a data sample of 300 tweets, of which a ratio of 80:20 has been used for training resp. testing. A 10k cross-validation has been used to ensure the validity of the test results.

The classifier has been programmed in Python. The Python package Scikit-learn was used for this study. Scikit-learn contains efficient and straightforward tools for data mining and data analysis and is open source. This package was selected because it contains multiple implementations of the Naïve Bayes algorithm. From the literature, it was seen that this is one of the more popular methods for test classification. However, the package was also chosen because it has an variety of different supervised Machine Learning algorithms, which allows us to investigate and compare different algorithms. We have used the following algorithms: Linear Support Vector Machines (LSVM), Naïve Bayes (NB), Decision Trees (DT), ExtraTrees (ET), Stochastic Gradient Descent (SGD) and Random Forests (RF). Of Naïve Bayes, three different implementations have been used: Gaussian Naïve Bayes (G-NB), Bernoulli Naïve Bayes (B-NB) and Multinomial Naïve Bayes (M-NB).

The data is labelled in two different classes. The first class are true messages. These messages are correct messages sent by humans. When detected, the classifier will label these messages with the integer '1'. The other class are false messages. These messages are sent by bots or are incorrect and misleading. When detected, the classifier will label these messages with the integer '0'.

Each algorithm has been trained and validated using the 10-fold cross-validation method. The best performing algorithm will be used to label the dataset. The performance of each Machine Learning algorithm can be found in section 4.

3.3 Qualitative content analysis

The selected Machine Learning algorithm labelled 613.033 Dutch tweets send between August 23 and November 1, 2012.

Table 1. Classification Report: Precision (P), Recall (R) and F-Score (F). Linear Support Vector Machines (LSVM), Gaussian Naïve Bayes (G-NB), Bernoulli Naïve Bayes(B-NB) and Multinomial Naïve Bayes(M-NB), Decision Trees (DT), ExtraTrees (ET), Stochastic Gradient Descent (SGD) and Random Forests (RF)

	False (0)			*True (1)*			*Mean*		
	P	R	F	P	R	F	P	R	F
LSVM	0,95	0,75	0,84	0,82	0,97	0,89	0,**88**	0,87	0,86
B-NB	0,66	0,89	0,76	0,86	0,59	0,7	0,77	0,73	0,73
M-NB	0,79	0,82	0,81	0,84	0,81	0,83	0,82	0,82	0,82
G-NB	0,88	0,79	0,83	0,83	0,91	0,87	0,85	0,85	0,85
DT	0,86	0,89	0,88	0,9	0,88	0,89	0,**88**	0,**88**	**0,88**
RF	0,69	0,86	0,76	0,84	0,66	0,74	0,77	0,75	0,75
ET	0,77	0,82	0,79	0,83	0,78	0,81	0,8	0,8	0,8
SGD	0,63	0,86	0,73	0,82	0,56	0,67	0,73	0,7	0,69

The algorithm detected messages that have possible false content sent intentionally or tweets posted by non-human users. A qualitative content analysis is conducted on these tweets, trying to find different patterns in the messages and the corresponding accounts. We categorised the detected false messages into different groups. These groups are created by the researcher. Results of this analysis can be found in section 4.2.

4. RESULTS

In this section, we will present the outcomes of our research. In section 4.1 we will show a classification report and a confusion matrix. Additionally, we will compare the different algorithms used and select one of these algorithms to use for the complete dataset. In section 4.2, we will analyse the Twitter dataset labelled by our Machine Learning algorithm.

4.1 Results of Classification

Performance measures have been used to evaluate the different algorithms. A confusion matrix and a classification report have been made, which can be found in Table 1 and 2.

Table 2. Confusion Matrix. Abbreviations for algorithms are like in Table 1.

		False (0)	*True (1)*
False (0)	LSVM	21	7
	B-NB	25	3
	M-NB	23	5
	G-NB	22	6
	DT	25	3
	RF	24	4
	ET	23	5
	SGD	24	4
True (1)	LSVM	21	7
	B-NB	25	3
	M-NB	23	5
	G-NB	22	6
	DT	25	3
	RF	24	4
	ET	23	5
	SGD	24	4

To determine which algorithms are best for this dataset, we look at the weighted F-Score. The F-Score is the harmonic mean of the precision rate and the recall rate. Precision is the ratio of true positives to all positives, while recall is the ratio of true positives to all correctly classified messages [30]. Three algorithms scored a high F-Score. The Decision Tree is the best performing with a weighted F-Score of 88%. Close to the Decision Tree are the Linear Support Vector Machine algorithm with 86% and the Gaussian Naïve Bayes algorithm with 85%.

Also, the low performance of the Random Forest was unexpected. The algorithm scored average with a weighted F-Score of 75% and was beaten by the Gaussian Naïve Bayes. With Random Forest each tree in the ensemble is built from a sample. When splitting a node during the construction, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases, which is generally compensated due to averaging [25]. For our dataset, this was apparently not compensated enough.

The Confusion Matrix shows that the Decision Tree algorithm and the Bernoulli Naïve Bayes have the highest score with the detection of fake messages. Both have the least false negatives. However, the Bernoulli Naïve Bayes has a really high False Positive rate. 41% of the real messages are classified as false by the algorithm. Due to this, the Bernoulli has the lowest weighted F-Score of 73%.

When looking at the messages that are classified as real, we can see that the Linear Support Vector Machine algorithm had the best performance, with only one message that is classified as false positive. Also, the Gaussian Naïve Bayes and the Decision Tree algorithm had a meagre false positive rate. LSVM had a false positive rate of 3%, while G-NB had a rate of 9% and DT 12,5%.

When comparing these three algorithms, it can be seen that the Decision Tree algorithm is the algorithm that suits our dataset the best. Despite the G-NB and the LSVM having both a high F-Score, relatively seen they are not very good in determining which messages are fake. The DT algorithm is best for classifing both true and false messages, hence the highest F-Score.

4.2 Results of Qualitative content analysis

A Decision Tree algorithm has been executed to detect false messages. Through the qualitative content analysis of false

tweets sent during the election, distinctive features and characteristics of false content have been found and grouped into six different categories. The results of this classification will be analysed in this section.

In total, 613.033 tweets have been categorizing by the Machine Learning algorithm. Of these tweets, 328.897 tweets have been identified as true. The rest of the tweets, 284.136, have been identified as false. Qualitative content analysis has been performed on the detected fake messages. The tweets are examined to find distinctive features and characteristics. It was investigated to what purpose the messages are sent. Distinctive characteristics have been found and grouped together to form six different categories. The different categories are:

- Satirical messages
- Sales messages
- Link-only messages
- Hashtag-only messages
- Retweets
- Negative persuasion

In the next sections, all categories will be explained. Also, an example of a tweet of the corresponding category will be given.

4.2.1 Satirical messages

The first category is satirical messages. These messages are untrue, and this is generally known by the receivers. They are put online intentionally but do not want to have a direct influence on elections or general opinion of voters. These tweets are put online for the entertainment of the receiver. Satirical Twitter accounts found in the database are impersonating famous people or acting like a news station. For instance, a Twitter account sends untrue messages while pretending to be the Queen of The Netherlands.

Example of tweet: "*Peiling: 80 % van zwevende kiezers inmiddels slapend. #slotdebat #nosdebat"*

4.2.2 Sales messages

Another category is sales messages. These messages used the Dutch election to advertise their products. With the use of popular hashtags and statements about hot topics, people were persuaded to visit companies websites on which products or services could be bought. However, the messages do not have any link with the products sold on the website. The hashtags of political parties or election topics are used for the distribution of the advertisements, not to share news, facts or opinions about politics.

Example of tweet: "*Stem u op #PvdD? Kijk dan bij de #Ergotherapeut - #tk2012 - http://t.co/YnNrwkiW"*

4.2.3 Link-only messages

The dataset contained much persuasive tweets links. The tweets consist of a hashtag with a link. All messages are sent by different accounts. These accounts tweet a lot of the same messages in a short amount of time, before the account is suspended. The accounts that tweeted such messages that are in our database were online for at most three days before getting suspended. According to the Twitter guidelines, tweeting messages only containing links without an description is considered spam [8].

Most links in these tweets are suspended. The links of a lot of these tweets found in our database refer to websites hosted on the tk domain, a top-level domain of Tokelau. These domains are presumably used because these domains are free of charge, so no payment details are needed. Therefor, it is hard to trace the original sender of these tweets.

Example of tweet: "*#Rutte http://t.co/qjA6dFb8"*

4.2.4 Hashtag-only messages

A lot of tweets have been sent containing only a hashtag of the abbreviation of the party. Most of the accounts that have to send these sorts of tweets are suspended. These tweets can be used to enlarge the name recognition and awareness of the political party.

Example of tweet: "*#VVD"*

4.2.5 Retweets

Bots have been used to retweet posts of regular people. These retweet bots can be used for multiple different goals. Having a high number of followers and a high number of retweets on your tweets can boost your image. When a message has a high number of retweets, it looks as if a lot of people agree with the message of the tweet. It makes the sender look more professional about the topic he was addressing (Marrs, 2018).

Unfortunately, retweets are hard to recognize by the Machine Learning algorithm, since they do not contain other words than the original message. By only using the content of the tweet, it cannot be determined whether the account that sends the retweet is a bot or not. More information about the account that retweeted the message is needed to determine if the sender is a bot.

Example of tweet: "*RT @alemanzio: Wat gaan we stemmen en waarom? ik blijf neutraal en de beste zinnen worden gepromoot. Help ze in den haag #stemmen"*

4.2.6 Negative persuasion messages

Fake messages have been used to negative persuade voters to not vote for a particular party. This has been done using false claims about a particular party, which can affect the opinion of a voter about a specific party. The sender tries to convince the receiver not to vote for a specific person or party in election time, or to damage a person. These messages often using sensationalist, dishonest, or outright fabricated headlines to increase readership, online sharing, and internet click revenue (Hunt, 2016)

Example of tweet: *"#PvdA en Stelende leden lijken onlosmakelijk met elkaar verbonden"*

5. CONCLUSION AND FURTHER WORK

In this research, we studied the how fake messages are used on Twitter during the Dutch election of 2012. A classifier was developed to detect potential fake messages. A dataset of Twitter messages about the Dutch election of 2012 has been classified. A qualitative content analysis has been performed on these classified tweets. In this chapter, the results of our research will be discussed, even as the limitations of the research and possible further work.

We analysed and compared eight different supervised Machine Learning algorithms. These algorithms have been trained and tested on a data sample of Dutch tweets about the Dutch election of 2012. The sample contained an equal number of 'true' and 'false' tweets. The sample has been used to train multiple supervised Machine Learning algorithms. The performance of these algorithms has been compared using the F-Score. The best performing algorithm, the Decision Tree algorithm, was used to label 613.033 Dutch tweets sent during the election period of the Dutch election of 2012. These false-labelled tweets were analysed in detail. Distinctive features of

the different false tweets were found, and the tweets were categorized into six different categories.

As a result, we can conclude that the Decision Tree algorithm is the best algorithm for the classification of true and false messages. The algorithm performed best with a weighted F-Score of 88%. However, the performance of the Linear Support Vector Machine algorithm was also worth mentioning with a weighted F-Score of 86%. However, the computation of a Linear Support Vector Machine algorithm is computationally expensive and highly dependent on the size of input data set [25]. Since our research needed a classifier that can handle a considerable amount of data, this algorithm is less suitable for the research then the Decision Tree algorithm.

The Decision Tree algorithm was used to classify different tweets in the database into either true messages or false messages. In total, 613.033 tweets were classified, of which 328.897 were classified as true, and 284.136 tweets were classified as false. These messages have been analysed, and the false data was categorized into six different categories: satirical messages, sales messages, link-only messages, hashtag-only messages, retweets and negative persuasion. Unfortunately, not all 284.136 tweets were indeed false. When performing the qualitative content analysis, it was noted that the data had some false positive labels.

Our research shows limitations that can be addressed and improved in future research. We used a dataset that was used by a research performed by [14]. This dataset contained twitter messages of 2012. These messages are six years old. This meant that many tweets that were identified as false were not online anymore. The user accounts were deleted or suspended. This meant that information about the user, such as Twitter followers and number of sent messages, could not be retrieved. Therefore, the Machine Learning algorithm was trained on only the content of the twitter message. It could be interesting to investigate the detection of potential fake messages with a combination of both the content of the tweet and the account data of the user that tweeted the message. As researched by Camisani-Calzolari [4], potential fake messages can also be identified using multiple data of the account. An algorithm that is trained on a combination of the content of the tweet and the account data could have a higher validity.

Due to the limited information of the user account, it was hard to make a substantial training set for the training of the machine learning. A more extensive training set could improve the validity of the Machine Learning algorithm and therefore decrease the amount of false-positive classifications.

6. REFERENCES

[1] Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. The Journal of Economic Perspectives: A Journal of the American Economic Association, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

[2] Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and visualisation of misleading content on Twitter. International Journal of Multimedia Information Retrieval, 7(1), 71–86. https://doi.org/10.1007/s13735-017-0143-x

[3] Brownlee, J. (2017, September 29). How to Prepare Text Data for Machine Learning with scikit-learn. Retrieved June 24, 2018, from https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn

[4] Camisani-Calzolari, M. (2012). Analysis of Twitter followers of the US Presidential Election candidates: Barack Obama and Mitt Romney. http://digitalevaluations.com.

[5] Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media and Society, 16(2), 340–358. https://doi.org/10.1177/1461444813480466

[6] Craig Silverman, J. S.-V. (2016, December 7). Most Americans Who See Fake News Believe It, New Survey Says. Retrieved May 29, 2018, from https://www.buzzfeed.com/craigsilverman/fake-news-survey

[7] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. Decision Support Systems, 80, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

[8] De Twitter-regels. (n.d.). Retrieved July 1, 2018, from https://help.twitter.com/nl/rules-and-policies/twitter-rules

[9] Dudovskiy, J. (n.d.). Snowball sampling. Retrieved June 25, 2018, from https://research-methodology.net/sampling-in-primary-data-collection/snowball-sampling/

[10] Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences of the United States of America, 112(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

[11] Exposed: Undercover secrets of Trump's data firm. (2018, March 20). Retrieved May 4, 2018, from https://www.channel4.com/news/exposed-undercover-secrets-of-donald-trump-data-firm-cambridge-analytica

[12] Goodman, L. A. (1961). Snowball Sampling. Annals of Mathematical Statistics, 32(1), 148–170. Retrieved from http://www.jstor.org/stable/2237615

[13] Graham-Harrison, E., & Cadwalladr, C. (2018, March 21). Cambridge Analytica execs boast of role in getting Donald Trump elected. The Guardian. Retrieved from http://www.theguardian.com/uk-news/2018/mar/20/cambridge-analytica-execs-boast-of-role-in-getting-trump-elected

[14] Hosch-Dayican, B., Amrit, C., Aarts, K., & Dassen, A. (2014). How Do Online Citizens Persuade Fellow Voters? Using Twitter During the 2012 Dutch Parliamentary Election Campaign. Social Science Computer Review, 34(2), 135–152. https://doi.org/10.1177/0894439314558200

[15] How to use hashtags. (n.d.). Retrieved May 4, 2018, from https://help.twitter.com/en/using-twitter/how-to-use-hashtags

[16] Hunt, E. (2016, December 17). What is fake news? How to spot it and what you can do to stop it. The Guardian. Retrieved from http://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

[17] Keretna, S., Hossny, A., & Creighton, D. (2013). Recognising User Identity in Twitter Social Networks via Text Mining. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 3079–3082). https://doi.org/10.1109/SMC.2013.525

[18] Ingram, D. (2018, March 20). Factbox: Who is Cambridge Analytica and what did it do? Reuters. Retrieved from https://www.reuters.com/article/us-facebook-cambridge-analytica-factbox/factbox-who-is-cambridge-analytica-and-what-did-it-do-idUSKBN1GW07F

[19] Marrs, M. (2018, January 12). Buying Twitter Followers: The (Cheap) Price of Friendship. Retrieved July 1, 2018, from https://www.wordstream.com/blog/ws/2013/05/16/buying-twitter-followers-cheap-price-friendship

[20] Martin, D. (2018, March 27). What role did Cambridge Analytica play in the Brexit vote? | DW | 27.03.2018. Deutsche Welle. Retrieved from http://www.dw.com/en/what-role-did-cambridge-analytica-play-in-the-brexit-vote/a-43151460

[21] Massey, T. (2017, October). Analysing the trend of Islamophobia in the UK using Machine Learning and Trend Analysis (MSc). Cardiff University.

[22] Munoz, A. (2014). Machine learning and optimization. Courant Institute of Mathematical Sciences, New York, NY, 14. Retrieved from https://cims.nyu.edu/~munoz/files/ml_optimization.pdf

[23] Nolan, H. (2014, March 13). Twitter Is Public. Retrieved June 22, 2018, from http://gawker.com/twitter-is-public-1543016594

[24] Oosterveer, D. (2018, January 29). Social media in Nederland 2018: uittocht van jongeren op Facebook. Retrieved May 13, 2018, from https://www.marketingfacts.nl/berichten/jongeren-keren-facebook-massaal-de-rug-toe

[25] Pedregosa, et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research: JMLR, 12(Oct), 2825–2830. Retrieved from http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[26] Scott, M. (2018, March 27). Cambridge Analytica helped "cheat" Brexit vote and US election, claims whistleblower. Retrieved June 5, 2018, from https://www.politico.eu/article/cambridge-analytica-chris-wylie-brexit-trump-britain-data-protection-privacy-facebook/

[27] Spierings, N., & Jacobs, K. (2018). Political parties and social media campaigning. Acta Politica. https://doi.org/10.1057/s41269-018-0079-z

[28] Weaver, M. (2018, April 17). Cambridge Analytica: ex-director says firm pitched detailed strategy to Leave.EU. The Guardian. Retrieved from http://www.theguardian.com/uk-news/2018/apr/17/cambridge-analytica-brittany-kaiser-leave-eu-brexit

[29] Vanderschoot, K. (2018, March 27). Heeft Cambridge Analytica de Britten naar de brexit geduwd? Retrieved June 5, 2018, from https://www.vrt.be/vrtnws/nl/2018/03/27/heeft-cambridge-analytica-de-britten-naar-de-brexit-geduwd--en-n/

[30] What is an intuitive explanation of F-Score? - Quora. (n.d.). Retrieved July 1, 2018, from https://www.quora.com/What-is-an-intuitive-explanation-of-F-Score

7. APPENDIX: CAMISANI-CALZOLARI RULE SET

The Camisani-Calzolari rule set [4] can be found in Table 3.

Table 3. Camisani-Calzolari rule set [4].

1. the profile contains a name;	**12.** it has used a hashtag in at least one tweet;
2. the profile contains an image;	**13.** it has logged into Twitter using an iPhone;
3. the profile contains a physical address;	**14.** it has logged into Twitter using an Android device;
4. the profile contains a biography;	**15.** it is connected with Foursquare;
5. the account has at least 30 followers;	**16.** it is connected with Instagram;
6. it has been inserted in a list by other Twitter users;	**17.** it has logged into *twitter.com* website;
7. it has written at least 50 tweets;	**18.** it has written the userID of another user in at least one tweet, that is it posted a *@reply* or a *mention*;
8. the account has been geo-localized;	**19.** (2*number followers) $\geq$ (number of friends);
9. the profile contains a URL;	**20.** it publishes content which does not just contain URLs;
10. it has been included in another user's favorites;	**21.** at least one of its tweets has been *retwitted* by other accounts (worth 2 points);
11. it writes tweets that have punctuation;	**22.** it has logged into Twitter through different clients (worth 3 points).