Sexual health information: credible or not?

An exploratory research into the automatic credibility assessment of online sexual health information using textual markers.

Author:
Jo Schreurs
Utrecht University
Graduate School of Natural Sciences

Supervisors:
Mirjam Visscher
Utrecht University
Graduate School of Natural Sciences

Remco Veltkamp
Utrecht University
Department of Information and Computing Sciences

Annemarie van Oosten
University of Amsterdam
Faculty of Social and Behavioural Sciences

Tuesday 16th August, 2022
Master Thesis Applied Data Science

Abstract

Nowadays, youth use online media as their primary source of sexual health information. However, user-generated content is not always reliable and can cause health problems, especially since it is difficult to distinguish between credible and unreliable information. Multiple studies have been done on automatic credibility assessment of online media, but not specifically sexual health information. These studies make use of markers that may indicate misinformation. Therefore, this study aims to obtain an overview of these markers and to examine whether and how these markers can be applied to the automatic credibility evaluation of user-generated sexual health data.

Based on the literature, this study created a comprehensive overview of all content-based markers (i.e., markers that could be derived from the text) that aided in credibility detection. A subset of these markers was modelled on the data using both supervised machine learning and more conventional methods. Subsequently, their relationships were examined to see if they aligned with the literature.

This study illustrates the disagreement between the existing literature on how the markers aid in credibility detection. Besides, the results indicate that there were no relationships between the vast majority of markers. This may be the first indication that most markers from the literature cannot be generalised to the automatic credibility assessment of sexual health data. However, this cannot be said with certainty due to the lack of a labelled dataset. The study listed several ideas for future research if a labelled dataset became available, which emphasized the idea that content-based markers should be used in combination with other markers.

Contents

1 Introduction

Vast amounts of sexual health information can be found online nowadays, which young people often rely on to learn about sex. A large-scale study amongst youth in the Netherlands shows that young people use the internet as their primary source of information (De Graaf, Van den Borne, Nikkelen, Twisk, & Meijer, 2017). This is because it is easily accessible, quick, and anonymous, amongst other things (Doornwaard et al., 2017; Kanuga & Rosenfeld, 2004). Nikkelen, van Oosten, and van den Borne (2020) distinguish between two types of online sexual information: professional websites about sex and interactive user-generated content (UGC). The latter has the advantage that young people can actively participate by creating and responding, and learn about other people's experiences (Attwood, Barker, Boynton, & Hancock, 2015; Kanuga & Rosenfeld, 2004).

However, the danger of seeking sexual health information online is that youth can access this information without parental supervision, increasing the risk of viewing age-inappropriate content (Kanuga & Rosenfeld, 2004). Besides, UGC is not created by professionals (Chou, Prestin, Lyons, & Wen, 2013). Consequently, this information is often not scientifically proven or in line with clinical practice. This can lead to serious health problems (Zhao, Da, & Yan, 2021). Lastly, online (social) media are perfect for spreading rumours, fake news and misleading information due to their growing popularity (Zhang & Ghorbani, 2020). It is, therefore, difficult for users to distinguish between credible and non-credible information.

Additionally, Fogg and Tseng (1999) states that people give more credence to a computer product - in this case, the UGC - if it orients them in unfamiliar situations and when they have a strong need for information. Thus, when youth are unfamiliar with specific information on sexual health and want their questions answered, they are even more likely to perceive that the information they find online is credible.

For the reasons mentioned above, sexual health information must be carefully evaluated to ensure that it is trustworthy and not dangerous (Eysenbach, 2008). The manual evaluation of such information is, however, very time-consuming. While there has been much research on recognizing deceitful or credible information, rumours, trolls, and fake news in online media (e.g., Addawood, Badawy, Lerman, and Ferrara, 2019; Castillo, Mendoza, and Poblete, 2011; Janze and Risius, 2017;Olteanu, Peshterliev, Liu, and Aberer, 2013; Zhou, Burgoon, Nunamaker, and Twitchell, 2004), only little research has focused on the detection of *health* misinformation (e.g., Mukherjee, Weikum, and Danescu-Niculescu-Mizil, 2014; Zhao et al., 2021), and none specifically on *sexual* health misinformation. Accordingly, this research focuses on exploring the automation of the credibility evaluation of such user-generated sexual health information.

More specifically, these studies on the credibility of information in online media mainly focus on textual cues, i.e., markers, that can help detect misinformation. Therefore, this study aims to obtain an overview of these markers and to examine whether and how these markers can be applied to the automatic credibility evaluation of user-generated sexual health data.

This research is divided into two parts: a theoretical and empirical part. The theoretical part aims to provide an overview of the markers that are proven to aid in credibility evaluation, and in which way. The empirical part explains how the markers are modelled on the available data, and the relationship between the markers is examined. This part also investigates the automation of readily available markers that were manually coded on sexual health data by researchers from the University of Amsterdam using machine learning methods. If there are clear relations between the markers that align with the literature, this research could be a first step in automating the credibility assessment of sexual health data. On the other hand, if there is no clear relation between the markers, this can indicate that other options for assessing credible sexual health information should be explored.

This research is organized as follows. Chapter 2 focuses on the definition of credibility and provides an overview of the different markers that are useful in detecting credible information. Chapter 3 describes the data that is used for this research, and Chapter 4 discusses the methods for extracting the credibility markers from sexual health data automatically and comparing the relationships between the markers. Chapter 5 presents the results, and Chapter 6 summarizes the key findings and conclusions from this research, as well as limitations and ideas for possible future research in the credibility assessment of sexual health information.

2 Literature overview

2.1 Credibility

This research focuses on detecting credible information. There are many definitions of credibility. The Merriam-Webster dictionary defines credible as "offering reasonable grounds for being believed" (Merriam-Webster, n.d.). One can thus think of credibility in terms of believability. According to Fogg and Tseng (1999), believability is also often used as a synonym for credibility, as well as "trust the information," and "accept the advice." They even divide credibility into two key components, namely expertise and trustworthiness. Together they make up how credible information is.

Furthermore, Wawer, Nielek, and Wierzbicki (2014) state: "Credibility, which is partly subjective, should not be confused with truth – usually understood as an objective category." Credible information should thus not always have to be factually accurate, but it is most likely to be believable. This research will also adhere to this definition of credibility.

Other studies specifically into the credibility of online information use other words instead of credibility, such as misinformation, deception, fake news, or rumours (e.g., Janze and Risius, 2017; Zhao et al., 2021; Zhou et al., 2004, and Zubiaga, Liakata, and Procter, 2017. Although these words are not always the exact synonyms of credibility, most articles do use a lot of overlapping markers in the detection of credible information. Therefore, this research still considers those articles that do not mention credibility to create an overview of the credibility markers.

2.2 Markers

Bondielli and Marcelloni (2019) conducted a survey to provide a comprehensive overview of rumour and fake news detection methods in recent literature. They found that most authors differentiate between *content*- and *context*-based features (i.e., markers). Zhang and Ghorbani (2020) also made a distinction between these categories. Their aim was, amongst other things, to give an overview of the state-of-the-art detection methods for fake news. They also distinguished a third category: *creator/user*-based features that "aim to capture the unique characteristics of suspicious user accounts" (Zhang & Ghorbani, 2020). Since the definitions of the categories are not quite the same, the categories and definitions of Bondielli and Marcelloni (2019) will be adhered to for this study. The reason for this is that this research mainly focuses on linguistic markers, which the study of Bondielli and Marcelloni (2019) describes in greater depth.

According to Bondielli and Marcelloni (2019), content-based features can be directly extracted from text, whereas context-based features are concerned with surrounding information. This research will focus on content-based features, as the available dataset only contains written messages and thus does not allow for a focus on context-based features (see Section 3). Within content-based features, Bondielli and Marcelloni (2019) distinguish between lexical, syntactic, and semantic linguistic markers. This article discusses many example markers for each of these categories. However, to keep it clear, only those markers that were proven to aid in the detect deception in at least one study will be discussed in more detail. Besides, at least one study must have mentioned how it aided in the detection of credible information.

2.2.1 Lexical markers

Lexical markers involve the actual word usage in texts, like swear words and self-reference. Table 1 provides an overview of all the lexical markers and how they aid in determining credibility according to different studies. Research from Addawood et al. (2019) on identifying political trolls on social media via linguistic cues found that trolls are less likely to use first-person singular pronouns like 'I', 'me', or 'my'. In other words, trolls are less likely to write from a personal perspective. The reason for the little use of self-reference could be that trolls do not want to take ownership of a statement. Zhou et al. (2004) also found that deceitful senders use less self-reference and more non-immediate and vague language. Interestingly, personal perspective or self-reference was also already manually coded as a marker for credibility by coders researchers from the University of Amsterdam in a dataset that will be used in this research (see Chapter 3.2).

Additionally, various articles researched the use of group reference by deceitful senders (Addawood et al., 2019; Zhou et al., 2004). Addawood et al. (2019) measured group reference through the use of third-person pronouns like 'they' and 'she' and found that deceitful senders used more group references. However, Zhou et al. (2004) defined this as 'other references' and found that these were ineffective in detecting deception. Zhou et al. (2004) measured group reference through the presence of first-person plural pronouns (i.e., 'we' and 'us') and concluded that deceivers use more group references.

Table 1: Overview of relevant lexical markers from literature.

Marker	Definition	Usage, Credibility	Author(s)
Personal perspective	Usage of first-person singular pronouns like 'I', 'me', or 'my'.	high, high	Addawood et al. (2019)
		low, low	Zhou et al. (2004)
Group reference	Usage of first-person plural pronouns like 'we', and 'us'.	high, low	Zhou et al. (2004)
Other reference	Usage of third-person pronouns like 'she', and 'they'.	high, low	Addawood et al. (2019)
		not effective	Zhou et al. (2004)

2.2.2 Syntactic markers

Table 2 provides an overview of all syntactic markers discussed in this section. Examples of syntactic markers are the number of content words, i.e., nouns, verbs, and adjectives (Bondielli & Marcelloni, 2019). Two studies from Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003) and Zhou et al. (2004) also found this. They concluded that the number of words, verbs, modifiers (i.e., adjectives, and adverbs), and noun phrases could aid in the detection of deception in texts. Other syntactic markers that they found were pausality (see equation 1) and lexical diversity (i.e., percentage of unique words in all words) (see equation 2). Agichtein, Castillo, Donato, Gionis, and Mishne (2008) trained two models to determine the answer and question quality separately. In this research, the unique number of words (i.e., lexical diversity) also proved to be a predictor of high quality answers, and the pausality – which they defined as punctuation density – was found to be a predictor for question quality.

$$pausality = \frac{\#\ of\ punctuation\ marks}{\#\ of\ sentences} \tag{1}$$

$$lexical\ diversity = \frac{\#\ of\ unique\ words}{\#\ of\ words} \tag{2}$$

Zhou et al. (2003) and Zhou et al. (2004) found that deceptive senders of texts use more words, verbs, modifiers, and noun phrases. Besides that, they display less lexical diversity and pausality (Zhou et al., 2004, 2003). Addawood et al. (2019) also found that a higher amount of words is a predictor of trolls on social media. However, their research showed that trolls use fewer verbs and modifiers. The research by Janze and Risius (2017), which focused on the automatic detection of fake news on social media platforms, likewise revealed that a higher word count increased the chance of fake news. The study of Zhao et al. (2021) aimed to build an effective model that could automatically detect health misinformation in online health communities. This study also found that a higher word count indicates misinformation.

Zubiaga et al. (2017) looked at improving a current state-of-the-art rumour detection system. They found that the current classifier's performance improved when it also relied on content-based features, which included the word count. Furthermore, Agichtein et al. (2008) concluded that the answer length is a good predictor for an answer of higher quality as well. However, both studies did not mention how this marker would help detect rumours. Lastly, the study of Alrubaian, Al-Qurishi, Hassan, and Alamri (2016) − aiming to create a credibility analysis system for information on Twitter − revealed that the difference in word count between credible and non-credible tweets is negligible.

Furthermore, Zhou et al. (2003) concluded that deceptive senders produce a higher typo ratio. This way, they appeared more informal than truth-tellers (Zhou et al., 2004). In their research on user perceptions of tweet credibility, Morris, Counts, Roseway, Hoff, and Schwarz (2012) found that non-standard use of grammar and punctuation indicates low perceptions of credibility. However, they noted that non-standard grammar might increase credibility in some user communities (Morris et al., 2012). Additionally, research from Jiaying, Song, and Zhang (2021) revealed that participants interpreted the same linguistic features differently for different information sources. They are less likely to take the number of typos into account during credibility judgement of online health information on forum web pages, compared to, for example, commercial websites. The research of Schwarz and Morris (2011), which aimed to establish a set of features that could help people more accurately judge the credibility of online content, omitted the number of spelling errors as a feature due to low correlation with credibility.

In addition, using specific punctuation can also play a role in detecting (non-)credible information. Zubiaga et al. (2017) used the use of a question mark, an exclamation mark, and a period as markers. According to their research, a question mark could be indicative of uncertainty. However, they did not mention specifically how the presence of exclamation and question marks and periods aided in detecting rumours. The study of Alrubaian et al. (2016) showed that almost all non-credible tweets in their dataset contained at least one question mark, whereas this is considerably less the case for credible tweets. The same study revealed that almost 99 per cent of credible tweets in their dataset did not contain any exclamation marks, while 80 per cent of non-credible tweets contained at least one (Alrubaian et al., 2016).

Addawood et al. (2019) likewise found that the use of question marks was significantly higher for trolls than non-trolls. Castillo et al. (2011) aimed to determine if the level of credibility of content posted on Twitter could be automatically assessed. They divided the tweets into topics and found that the fraction of Tweets containing question marks per topic related to low credibility. In contrast, the study of Janze and Risius (2017) revealed that whether a Facebook post or the post's title contained a question mark was ineffective in detecting fake news.

Olteanu et al. (2013) aimed to automate the credibility evaluation of a web page by only using information available on the web. They filtered out statistically irrelevant features for the credibility prediction task via feature selection. The results showed that the number of question marks in the text was a good indicator of credibility, but this was not the case for exclamation marks. According to their findings, exclamation marks had no particular effect on the credibility of web pages.

Other punctuation marks that could help detect credible information are quotation signs (Addawood et al., 2019; Janze & Risius, 2017). The study of Janze and Risius (2017) found that a Facebook post containing a citation – indicated by at least two quotation signs – was less likely to be fake news. In contrast, Addawood et al. (2019) concluded that the use of quotations was higher in trolls since it indicated uncertainty.

Table 2: Overview of relevant syntactic markers from literature.

Marker	**Definition**	**Usage, Credibility**	**Author(s)**
# of words	Total amount of words.	high, low	Addawood et al. (2019)
		not mentioned	Agichtein et al. (2008)
		not effective	Alrubaian et al. (2016)
		high, low	Janze and Risius (2017)
		high, low	Zhao et al. (2021)
		high, low	Zhou et al. (2003)
		high, low	Zhou et al. (2004)
		not mentioned	Zubiaga et al. (2017)
# of verbs	Total amount of verbs.	low, low	Addawood et al. (2019)
		high, low	Zhou et al. (2003)
		high, low	Zhou et al. (2004)
Typo ratio	$\frac{\#\ of\ misspelled\ words}{\#\ of\ words}$	not effective	Schwarz and Morris (2011)
		high, low	Zhou et al. (2003)
		high, low	Zhou et al. (2004)
Pausality	$\frac{\#\ of\ punctuation\ marks}{\#\ of\ sentences}$	high, low	Addawood et al. (2019)
		not mentioned	Agichtein et al. (2008)
		low, low	Zhou et al. (2003)
		low, low	Zhou et al. (2004)
Lexical diversity	$\frac{\#\ of\ unique\ words}{\#\ of\ words}$	not mentioned	Agichtein et al. (2008)
		low, low	Zhou et al. (2003)
		low, low	Zhou et al. (2004)
# of question marks	Total amount of question marks.	high, low	Addawood et al. (2019)
		yes, low	Alrubaian et al. (2016)
		yes, low	Castillo et al. (2011)
		not effective	Janze and Risius (2017)
		not mentioned	Olteanu et al. (2013)
		not mentioned	Zubiaga et al. (2017)
# of exclamation marks	Total amount of exclamation marks.	yes, low	Alrubaian et al. (2016)
		not effective	Olteanu et al. (2013)
		not mentioned	Zubiaga et al. (2017)
# of quotation marks	Total amount of quotation marks.	yes, high	Janze and Risius (2017)
		high, low	Addawood et al. (2019)

2.2.3 Semantic markers

The last type of markers found by Bondielli and Marcelloni (2019) are semantics, which are often extracted using more advanced natural language processing techniques. An example of a semantic marker that has proven to be useful in detecting fake news and rumours are sentiment analysis.

Sentiment analysis determines if a text is emotionally negative, positive, or neutral (Jain, Pamula, & Srivastava, 2021). The result of this analysis can also be referred to as polarity and is represented as a value between -1 and 1, where -1 is an indication for a negative, and 1 for a positive sentiment.

Castillo et al. (2011) found in their research that a higher fraction of tweets with a negative sentiment score meant higher credibility, and a low fraction of tweets with a positive sentiment score meant low credibility.

Alrubaian et al. (2016) determined the sentiment of a tweet by counting the number of negative and positive words that were based on a predefined list of sentiment words. The results showed that credible tweets tend to be more positive than non-credible tweets, and non-credible tweets are more likely to be negative than credible tweets.

Through sentiment analysis, the research of Zhao et al. (2021) found that misinformation was on average more positive than legitimate information. However, both the legitimate and misinformation in the dataset of this research contained more positive than negative information.

Table 3: Overview of relevant semantic markers from literature.

Marker	Definition	Usage, Credibility	Author(s)
Positive sentiment	Positive emotions, indicated by a number between 0 and 1.	high, high low, low high, low	Alrubaian et al. (2016) Castillo et al. (2011) Zhao et al. (2021)
Negative sentiment	Negative emotions, indicated by a number between -1 and 0.	high, high high, low	Castillo et al. (2011) Alrubaian et al. (2016)

2.2.4 Conclusion

In summary, one can distinguish between content- and context-based markers. Only the relevant content-based markers, which were proven significant in detecting credible information by earlier research, were discussed. However, there is often no consensus on how a marker impacts credibility. Nonetheless, to the author's knowledge, this is the first research in which an overview of the markers also mentioned the markers' impact on credibility. Besides, due to the exploratory nature of this research, it is not necessarily a problem that not all literature agrees. Suppose the relationships between the markers automated in this research align with most of the literature. In that case, this research can provide additional support for already existing literature. If contrary or no relationships are discovered, this may indicate that the markers are not generalizable to different data types, such as sexual health data. Since there are many markers, only a subset was used in this research for further investigation.

1. Word count
2. Personal perspective
3. Pausality
4. Lexical diversity

5. Number of question marks
6. Number of exclamation marks
7. Sentiment score

3 Data

Two datasets were used for this research. These datasets were both created and provided by a research team of the Faculty of Communication Science at the University of Amsterdam.

3.1 Dataset 1: FOK!forum comments

The first dataset used was `20210520_comment_list.csv`. It was created by scraping comments regarding the topic of sexual health between April 2016 and May 2021 from the website FOK!forum. This is a Dutch forum website which is part of the website FOK!. FOK! has an online community and contains mostly content such as news, reviews, columns, and opinion polls (Wikipedia, n.d.). All sorts of topics are discussed on the forum, including sports, media, hobbies, and sexual health. FOK!Forum is an unmoderated website, meaning that the website does not rely on the moderation of messages to minimize harmful messages (Wise, Hamman, & Thorson, 2006).

3.1.1 Data preparation

This dataset consisted of four columns (see Figure 1), namely `TopicID`, `CommentID`, `Comment`, and `Time`, where the column name `TopicID` was changed to `ThreadID` in consultation with the author of the dataset, since this was more fitting to the content of the column. This column contained the unique identifier of the thread under which the comment was posted. `CommentID` was the unique identifier of the comment itself and also the candidate key of this dataset, meaning that this column was needed to uniquely identify each row in the dataset (Silberschatz, Korth, & Sudarshan, 2002). The column `Comment` contained the actual text of a comment, and `Time` contained the day, date, and time of when a comment was posted on FOK!forum.

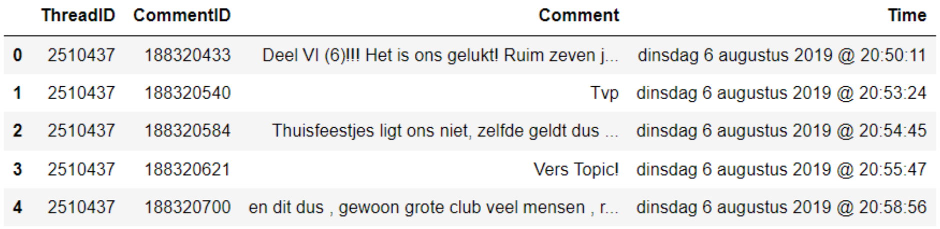
Besides that, the dataset contained 116962 rows. However, all the rows in which the column `Comment` contained missing values were deleted, which led to a final dataset of 113526 rows. The reason for deleting these missing values was that the markers used for this research were all content-based (see Chapter 2.2), i.e., the markers could be directly extracted from text (Bondielli & Marcelloni, 2019). Since these rows contained no comment and thus no text, they could not be used to identify markers.

3.2 Dataset 2: Manually coded markers

The second dataset, `Training dataset.csv`, contained manually coded markers identified by a research team from the University of Amsterdam. These markers also contribute to detecting credibility.

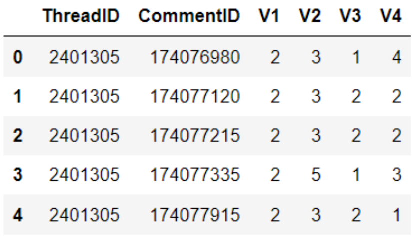
3.2.1 Data preparation

The dataset consisted of eight columns (see Figure 2). The first two columns, `Topic` and `rank` were not relevant to answer the research question. `ThreadID` was the unique identifier of a thread, and `Comment`



	ThreadID	CommentID	Comment	Time
0	2510437	188320433	Deel VI (6)!!! Het is ons gelukt! Ruim zeven j...	dinsdag 6 augustus 2019 @ 20:50:11
1	2510437	188320540	Tvp	dinsdag 6 augustus 2019 @ 20:53:24
2	2510437	188320584	Thuisfeestjes ligt ons niet, zelfde geldt dus ...	dinsdag 6 augustus 2019 @ 20:54:45
3	2510437	188320621	Vers Topic!	dinsdag 6 augustus 2019 @ 20:55:47
4	2510437	188320700	en dit dus , gewoon grote club veel mensen , r...	dinsdag 6 augustus 2019 @ 20:58:56

Figure 1: The first five rows of the first dataset containing the FOK!forum comments.



	ThreadID	CommentID	V1	V2	V3	V4
0	2401305	174076980	2	3	1	4
1	2401305	174077120	2	3	2	2
2	2401305	174077215	2	3	2	2
3	2401305	174077335	2	5	1	3
4	2401305	174077915	2	3	2	1

Figure 2: The first five rows of the second dataset containing the manually coded markers.

was the unique identifier of the comment itself. These columns contained the same information as `ThreadID` and `CommentID` in the first dataset (see Chapter 3.1.1). The dataset contained four columns with four manually coded different markers, of which only column `V3` was used for this research. This dichotomous marker indicates the presence of a personal perspective (e.g., by using words like 'I' and 'me') in the comment, and is thus comparable to the personal perspective marker in Chapter 2.2.1. To not confuse the two markers, the marker personal perspective from this manually coded dataset will be called `V3` personal perspective throughout the remainder of this paper.

The number of rows of this dataset was 9151. However, the columns `V3` contained a lot of missing values. The reason was that the coders were told to skip the comments that contained no information about sex and sexual behaviour. Therefore, these missing values were deleted, which led to a final dataset of 4061 rows. The percentage of comments per class ('Yes'/'No') of marker `V3` is visible in Figure 3. This graph shows slightly imbalanced data, with 1702 comments for class 'No' and 2359 for class 'Yes'.

3.3 Ethical and legal considerations

Using data created by people on the internet raises ethical and legal questions. The main question is if it is permitted to use this data for research purposes. There are several considerations to be made here.

First, the authors of the comments did not agree to be part of the first dataset mentioned above. However, FOK!forum is a public website. Therefore, users consented to publicly posting their comments on the internet, making them accessible for anyone to read and use. Besides, the author of this dataset did receive passive consent from FOK!forum to use their data. That is, she contacted FOK!forum and did not receive a response. Furthermore, their terms of use do not state that their data cannot be used for other purposes.

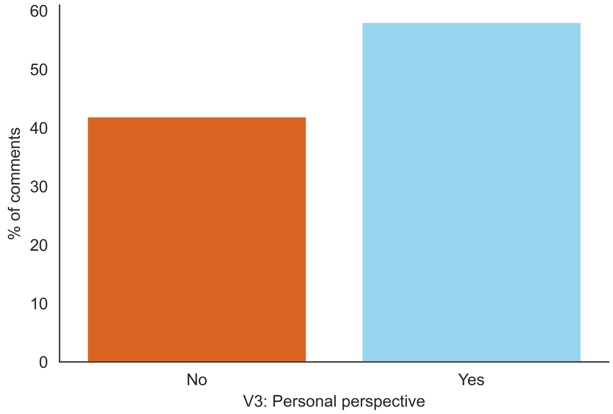


Figure 3: Percentage of comments per category of marker V3: personal perspective.

Furthermore, for the creation of this dataset, no usernames were scraped. Additionally, most usernames on FOK!forum do not reveal the identity of the users: most usernames only include the first name, or they do not include the user's name at all. The scraped data was thus anonymous; this way, potential harm to users was kept to an absolute minimum. The second dataset only included IDs of comments and threads. Since the actual text of the comments was not included, this dataset was anonymous as well, and there was no potential harm to users.

Lastly, a considerable advantage of using this data for research is that it provides the opportunity to properly educate youth about the quality and credibility of sexual health information. In conclusion, since the potential harm to users is minimal, the advantage outweighs the disadvantage of using this data without the user's explicit consent.

4 Method

The objective of this study was to examine whether and how the markers mentioned in Chapter 2.2 could be applied to the automatic credibility evaluation of user-generated sexual health data. In order to achieve this goal, several steps were taken.

First, supervised machine learning methods were used to see if the manually coded marker `V3` personal perspective could be predicted on the comments in the FOK!forum dataset that were not yet coded (see Chapter 4.1). Then the other markers were identified on the whole dataset, which is explained in Chapter 4.2. Afterwards, the relationships between these markers were explored to see if they aligned with the findings from existing literature. The data analysis was executed in Python unless mentioned otherwise. The notebooks with code for this research are available on `GitHub`.

4.1 Predicting manually coded marker

The marker `V3` personal perspective was already manually coded on a subset of the FOK!forum data (see marker `V3` in Chapter 3.2). Therefore, supervised machine learning was applied to see if it was possible to predict the marker on the FOK!forum data that were not yet labelled. It was the only marker corresponding to the existing literature and for which a labelled dataset was available. Suppose `V3` personal perspective could be predicted with supervised machine learning methods. In that case, this could indicate that it would be worthwhile to use similar methods to model other manually coded markers on unseen data.

`V3` personal perspective is a categorical variable, and the prediction of the class label (in this case, 'Yes' or 'No') of such a variable is called classification (Han, Pei, & Kamber, 2012). In order to train a machine learning model, it is important to pre-process the data and apply feature extraction (Kowsari et al., 2019). These steps will be explained in the following section, after which an explanation of the different classification approaches and their evaluation metrics follows.

4.1.1 Preprocessing

Data preprocessing is an essential first step in working with data. It ensures the data quality and can thus help ensure that the data is accurate, consistent, complete, and easily interpretable (Han et al., 2012). Text data is especially prone to being noisy, which can influence the performance of machine learning algorithms (Kowsari et al., 2019). This section will explain the preprocessing steps for predicting marker `V3` personal perspective.

Tokenization The first step in the preprocessing of the comments was tokenization, which means breaking a stream of text into tokens (Christopher, Prabhakar, & Hinrich, 2008; Kowsari et al., 2019). Kowsari et al. (2019) defines a token as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing." In this case, this meant splitting the comments into meaningful words. For this, the natural language processing tool `spaCy` was used (Honnibal & Montani, 2017). The reason for using this was that it had an available pipeline for Dutch texts. Therefore, it was able to split Dutch texts into meaningful tokens.

Stop word removal Another preprocessing possibility is the removal of stop words. These words are often of little value in classification algorithms (Christopher et al., 2008; Kowsari et al., 2019). Examples of Dutch stop words are 'slechts', 'jou', 'ik', 'onder'. Since the list of stop words includes pronouns, like 'ik', 'jou', 'zij', they were not removed from the comments. These pronouns can contribute especially to the classification of marker `V3` personal perspective since they indicate the presence of a personal perspective in the text.

Noise removal Special characters and punctuation are often crucial for the human understanding of texts, but not so much for classification algorithms (Kowsari et al., 2019). Therefore, punctuation and special characters like emojis were removed since these do not indicate personal perspective.

Stemming and lemmatization Stemming and lemmatization are both techniques to "reduce inflectional forms and sometimes derivationally related forms of a word to a common base form" (Christopher et al., 2008). However, these preprocessing techniques were both not used. The reason for this is that the form of a word is important for these markers. For example, the word 'had' indicates that the sentence is written in first person, whereas 'heeft' is an indication for a sentence written in third-person. When both words would have been transformed to 'hebben', this information would have been lost.

Capitalization Diverse capitalization can be a problem for the classification of large documents (Kowsari et al., 2019). A common strategy for solving this is case-folding, which reduces all letters to lower case (Christopher et al., 2008). Case-folding was applied as the last preprocessing step, and the reason for doing so was that it increased the model's performance.

4.1.2 Train-test set

After performing all preprocessing steps, the data was split into a train and test set. The training data size was 80 per cent and contained the data used to generate the classifier. The test set consisted of the remaining 20 per cent of the dataset. The function of the test set is to measure the performance of a classifier (Shami & Verhelst, 2007).

4.1.3 Vectorization

When using texts as input for machine learning algorithms, the texts must be vectorized, i.e., numerically represented. There are several ways to do this, of which $TF(t,d)$ is the most basic approach. $TF(t,d)$ stands for term frequency and corresponds to the number of times a term t occurs in a document d (scikit-learn, n.d.-b). In this case, a document is the same as a comment. The danger of this vectorization method is that it may cause commonly used words to dominate (Kowsari et al., 2019). For this reason, $TF - IDF(t,d)$ was calculated, which stands for term frequency-inverse document frequency (Kowsari et al., 2019). One can calculate it as follows:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

where $IDF(t)$ (inverse documents frequency) weighs how often a word occurs in other documents: frequently recurring words, like 'the', are given less weight, and terms that rarely occur get more weight (Spark Jones, 1972).$IDF(t)$ in is calculated using the following equation:

$$IDF(t) = log\frac{1+n}{1+df(t)} + 1 \tag{4}$$

where n is the total number of documents, and $df(t)$ is the number of documents containing term t (scikit-learn, n.d.-b). The matrix of each comment's TF-IDF scores per word was then used as input for the supervised machine learning algorithms.

4.1.4 Supervised machine learning algorithms

There are several possible machine learning approaches for classification problems. Two approaches were chosen and compared to find the most suitable model for predicting marker `V3` personal perspective on unseen data. Both models were trained using the training data and then tested with the test data. A brief explanation of the different approaches and the tuned hyperparameters follows.

Support Vector Machine Support vector machine (SVM) is a classification algorithm that aims to find an optimal line, i.e., decision boundary, to separate the data into classes (Han et al., 2012). This line is called a hyperplane. It does so by finding the support vectors, which are the nearest points to the decision boundary from each class. The distance between the line and the support vectors is computed, also known as the margin. The algorithm aims to maximize the margin between classes, and SVM thus searches for the hyperplane with the largest margin. However, often times data is not that easily linearly separable. Therefore, SVM finds the hyperplane by transforming the original training data into a linearly separable higher dimension. An advantage of SVM is that it is robust to overfitting (Kowsari et al., 2019). Due to the high-dimensional space, this is especially the case for text data.

The model `SVC()` from the `scikit-learn` package was used (Pedregosa et al., 2011). The first parameter that was tuned in this model was the kernel. Two kernels were chosen: `linear` and `RBF`. The latter is the standard non-linear kernel and thus allows for a non-linear decision boundary, meaning that no straight line would be able to separate the classes (Han et al., 2012).

Another important hyperparameter is the penalization factor (`C`), which controls the trade-off between the classification accuracy and complexity (scikit-learn, n.d.-a). Small values of `C` lead to lower complexity. However, this also increases the risk of data being misclassified. A higher `C` leads to higher classification accuracy.

The last hyperparameter that had to be tuned was `gamma`, which determines the influence of a single training example (scikit-learn, n.d.-a). When `gamma` becomes larger, other examples must be closer to be affected. It should be noted that `gamma` only has to be tuned when the kernel is non-linear; this hyperparameter can be discarded with a linear kernel.

In order to find the optimal kernel, value of `C` and `gamma` (in the case of a non-linear kernel being the most optimal), `GridSearchCV()` from `scikit-learn` was used in combination with 5-fold stratified cross-validation (Pedregosa et al., 2011). With `GridSearchCV()`, one can specify different values for each hyperparameter. `GridSearchCV()` then considers all parameter combinations and tests which combination leads to the best performing model (scikit-learn, n.d.-c).

Logistic Regression Another classification algorithm is logistic regression. This algorithm also aims to fit an S-shaped curve, i.e., a logistic curve, to the data in order to classify the observations (Singh, Thakur, & Sharma, 2016). It does so by calculating the probability of the input belonging to one of two classes (in this case, ‘Yes’ or ‘No’) (Nasteski, 2017). This probability is calculated for both classes, and the class with the highest probability is taken as the prediction.

A benefit of logistic regression is that it requires little to no tuning (Kowsari et al., 2019). However, there are still several parameters that can be tuned. Only one hyperparameter was chosen for tuning in this research, namely `C`. Like in SVMs, a smaller value of `C` leads to lower complexity and, thus, stronger regularization. Again, `GridSearchCV()` from `scikit-learn` was used in combination with 5-fold stratified cross-validation to find the optimal value for the hyperparameter `C`.

4.1.5 Evaluation measures

After the models were trained using the training set, the test set was used to test the classification performance of the models. The metrics used to evaluate and compare the performance were precision, recall, accuracy, and F1 score. All these measures have a score between 0 and 1. The closer the measure gets to 1, the better the model performs.

1. *Accuracy* is the number of correct predictions divided by the total number of predictions (see Equation 5) (Han et al., 2012; Silberschatz et al., 2002). The true positive (TP) rate is the number of correct positive predictions, and the false positive (FP) rate is the number of incorrect positive predictions. The true negative (TN) rate is the number of correct negative predictions, and the false negative (FN) rate is the number of incorrect negative predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

2. *Precision* indicates how often the positive predictions are correct; it is a measure of exactness (see Equation 6) (Han et al., 2012; Silberschatz et al., 2002).

$$precision = \frac{TP}{TP + FP} \tag{6}$$

3. *Recall* is a measure of completeness and indicates how many of the positive predictions are labelled as such (see Equation 7) (Han et al., 2012; Silberschatz et al., 2002).

$$recall = \frac{TP}{TP + FN} \tag{7}$$

4. The *F1-score* is the harmonic mean of precision and recall, and gives equal weight to both (see Equation 8) (Han et al., 2012; Silberschatz et al., 2002). Like accuracy, the F1-score quantifies the overall amount of error, but it is not as affected by imbalanced data.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

4.2 Extracting credibility markers

As mentioned in Chapter 2.2, seven credibility markers were chosen to be automated and compared. This section will explain how these markers were modelled on the data and which preprocessing steps were used. Table 4 provides an overview of the preprocessing steps per marker.

4.2.1 Number of words

To count the number of words, the text from the column `Comment` was tokenized using the `spaCy` package (Honnibal & Montani, 2017). This package was used since it includes a module to tokenize Dutch texts. Afterwards, all tokens that were not words (i.e., emojis, punctuation, and empty strings containing solely white spaces) were removed. The remaining tokens were then counted.

4.2.2 Lexical diversity

Lexical diversity was calculated using equation 2. For the calculation of the total number of words, one can refer back to 4.2.1. The unique number of words was calculated by lowercasing all the words whereafter all duplicate tokens per comment were removed to count the unique words.

4.2.3 Number of exclamation and question marks

Both the number of exclamation and question marks were calculated the same way. For this, the module `punctuation` from the `string` package was used. This module provides a list with all possible punctuation marks. With this list, the punctuation could be recognized and counted in the text.

4.2.4 Pausality

This `punctuation` module was also valuable for counting all punctuation in the text. This was necessary to calculate the pausality (see Chapter 1). Besides that, the total number of sentences per comment was calculated using the sentence tokenizer included in the `spaCy` package. This tokenizer could not perfectly tokenize all sentences. However, since the spaCy package is one of the few packages with a Dutch sentence tokenizer, this package provided the most accurate results.

4.2.5 Personal perspective

The presence of a personal perspective in a comment was found by searching for words related to self- or group reference. These words were ‘ik’, ‘me’, ‘mij’, ‘mijn’, ‘m’n’, and ‘mn’. When a comment included at least one of these words, it was considered to be written from a personal perspective.

This marker and the manually coded marker `V3` personal perspective both indicate a personal perspective in the text. Since these markers are the same except that one was automatically detected and one was manually detected, the inter-rater reliability was calculated to measure the degree of agreement between two raters: the computer and a human. In this research, Cohen’s kappa was used to calculate inter-rater reliability with `cohen_kappa_score()` from the `scikit-learn` package (Pedregosa et al., 2011). There are various kappa measures (e.g., weighted kappa, Fleiss’ kappa), but Cohen’s kappa is specifically used when there are two raters, the same two raters rate each subject, and all disagreements may be considered equally serious (Cohen, 1960; Fleiss, 1971). Cohen’s kappa tries to account for the fact that the agreement on some items may purely be due to chance and is calculated as follows:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{9}$$

where p_o is the proportion of units in which the annotators agreed, and p_e is the expected agreement when both annotators assign labels randomly (Cohen, 1960).

4.2.6 Sentiment

The sentiment score for the comments in the dataset was already calculated by Lai (2022). The two datasets were merged so the column `sentiment` could be added as a marker.

Table 4: Overview of the preprocessing steps as discussed in Chapters 4.1.1 4.2 and to which markers they were applied. Note: stop word removal and stemming/lemmatization are not included because they were used for the preprocessing of any of the markers.

	Tokenization	**Special characters and punctuation removal**	**Case-folding**
`V3` personal perspective	x	x	x
number of words	x	x	
personal perspective	x	x	x
lexical diversity	x	x	x
pausality	x		
# of question marks			
# of exclamation marks			

4.3 Exploring relations between markers

After the markers were detected, the relationship between the markers was assessed. Since there is no labelled dataset, it is difficult to determine if the markers were helpful in credibility detection. Examining the relationships can at least reveal whether the relations between the markers correspond to the literature. If so, this may be a first indication that the markers can be used to assess the credibility of sexual health information. Of course, follow-up research with a more extensive dataset would be necessary to confirm this. If there are contrary or no relationships, this may indicate that the markers cannot be generalized to sexual health data. This section explains how the relations between the markers were investigated.

4.3.1 Spearman's rank order correlation

To assess the relationship between the continuous markers, Spearman's rank order correlation was computed using `corr(method='spearman')` from the `pandas` package (McKinney, 2010). It can be calculated as follows:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{10}$$

where n is the total number of measured units (i.e., the sample size) and d^2 are the squared rank differences between variables (Astivia & Zumbo, 2017). Spearman's correlation coefficient determines the direction and strength of the monotonic relationship between two variables (Artusi, Verderio, & Marubini, 2002). In a monotonic relationship, two variables move in the same or opposite direction but not necessarily at the same rate.

Spearman's rank order correlation is a nonparametric version of the more commonly used Pearson's correlation coefficient (Xiao, Ye, Esteves, & Rong, 2016). However, Pearson's correlation coefficient assumes a linear relationship between the two variables and no extreme outliers (Burns & Burns, 2008). Both assumptions were violated, and thus this test was not applicable. In contrast, Spearman's rank order correlation assumes that the measured variables should be on an ordinal, interval or ratio scale, and the variables should represent paired observations (Burns & Burns, 2008). Both assumptions were satisfied for all markers except personal perspective, which was a nominal variable.

4.3.2 Box plots

In order to compare the relationship between the nominal variable personal perspective and the other continuous variables, the box plot of each continuous variable was plotted per class ('Yes' or 'No'), and the medians were compared.

No statistical test was used to investigate the relationship between the continuous and categorical variables because the data did not meet the assumptions for these tests. The first option was to calculate the point-biserial correlation coefficient. However, this test is sensitive to outliers and skewed distributions (Chao, 2017). All continuous markers had many outliers and their distributions were very skewed.

Another option would have been to perform logistic regression. However, this test assumes a linear relationship between each independent variable and the logit of the dependent variable (Stommel & Dontje, 2014). This assumption was not met. Another assumption was that the observations were independent of each other, which was not the case since many comments were in reply to each other. Thus one observation in the dataset influenced the other.

In conclusion, although several tests were considered, the assumptions of these tests were not met, which could make the results inconclusive. Therefore, comparing box plots seemed the most sensible option, as this gave a clear view of the median and the data distribution per class.

5 Results

This section discusses the results of the analyses explained in Chapter 4. First, the results of the classification of `V3` personal perspective using supervised machine learning models are discussed. Second, this Chapter explains the relations between the continuous markers. Third, the categorical marker personal perspective and continuous markers are compared using box plots. This last section also mentions the inter-rater reliability between the automatically coded marker personal perspective and the manually coded marker `V3` personal perspective.

5.1 Predicting `V3` personal perspective

Table 5 shows the classification results obtained with each machine learning model. The overall performance of both models is the same, with an accuracy of .90. This means that the models are able to accurately predict if there is a personal perspective or not in 90 out of 100 comments. Both models have the same precision for the positive class but score a little lower on the precision of the negative class. This means that both models were more often correct in classifying the positive ('Yes') than the negative class. However, recall – which is the ability to find all relevant instances of a class in the dataset – is higher in the positive class for logistic regression. In contrast, it is higher in the negative class for SVM. The F1-score, i.e., the weighted mean of precision and recall, is higher for logistic regression than SVM, but the difference is minimal. It also shows that both models are generally better at predicting the positive than the negative class. The time to train the models indicates that it took the SVM circa 58 times as much time to get trained as the logistic regression classifier.

Table 5: Classification results for logistic regression and support vector machine, as well as the time it took on average to train each classifier.

Model	Class	Precision	Recall	F1-score

Support Vector Machine
`C = 1`
`kernel='linear'`
	Yes	.93	.89	.91

	No	.84	.90	.87
	Accuracy	.90		
	Time	218.4s		
Logistic Regression `C=10`	Yes	.93	.91	.92
	No	.86	.89	.88
	Accuracy	.90		
	Time	3.74s		

The advantage of the best-performing model, logistic regression, is that it gives insight into which words are given the most weight in predicting personal perspective by computing the coefficients. The top fifteen coefficients with the largest absolute value were extracted and are visible in Figure 4. Positive coefficients indicate the most prevalent terms in comments with personal perspective, with the word 'ik' having by far the highest coefficient. On the other hand, the negative coefficients are associated with terms that belong to comments with no personal perspective, where the term 'je' has the lowest coefficient.

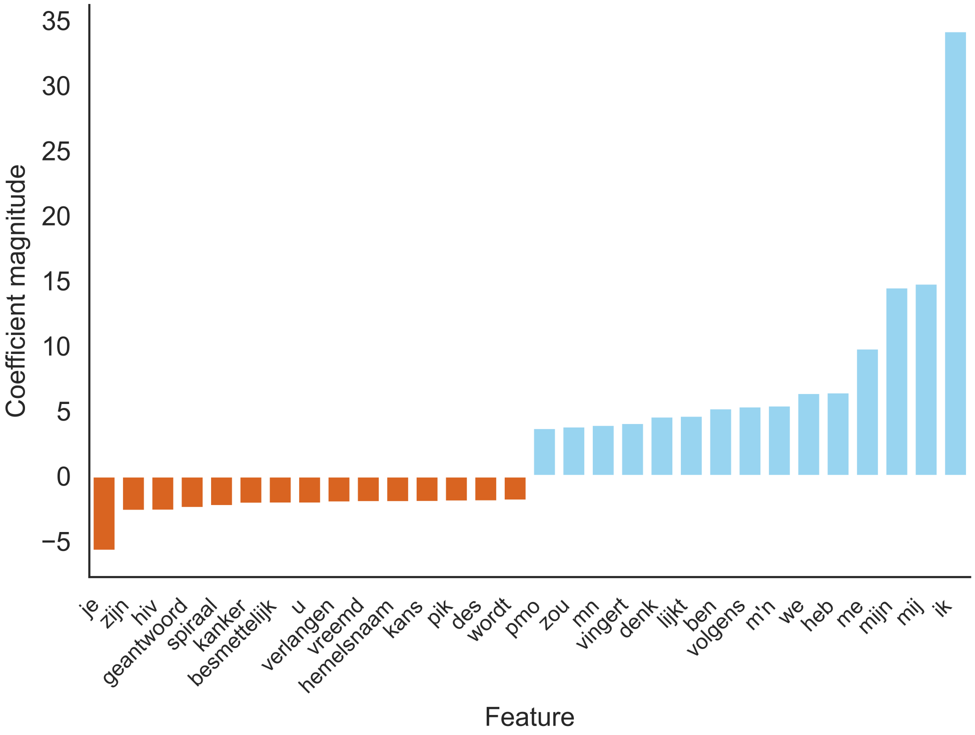


Figure 4: The fifteen coefficients that are given most weight in predicting `V3` personal perspective in the logistic regression model. Positive coefficients indicate terms that are important in predicting the positive class ('Yes'), whereas the negative coefficients are terms that are indicators of a comment belonging to the negative class ('No').

5.2 Comparing markers

This section discusses the relations between the markers that were automatically modelled on the data (see Chapter 4.2). First, the descriptive statistics of the continuous variables and their relations are evaluated. Second, the automatically coded marker personal perspective is compared to the continuous markers and the manually coded marker `V3` personal perspective.

5.2.1 Continuous variables

Descriptive statistics The descriptive statistics of the continuous markers are available in Table 6. The word count was generally quite low (*M*=31.71, *SD*=78.29). However, a visual inspection of the data showed that the data contained many outliers, with the maximum word count being 6468. Pausality was also generally quite low (*M*=.91, *SD*=.12) with a few extreme outliers. With the mean, median, first quartile and third quartile all being between one and two, it can be inferred that at least half of the comments contained between one and two punctuation marks per sentence.

Furthermore, the lexical diversity was very high (*M*=.96, *SD*=.12), meaning that the fraction of unique words per comment was high. The minimum was 0 since some comments did not contain any words but only special characters.

Both the number of question marks (*M*=.34, *SD*=.82) and the number of exclamation marks (*M*=.13, *SD*=.59) were generally very low, with at least half of the comments not containing any at all, indicated by a median of 0. A visual inspection did show a few extreme outliers, with a maximum of 77 for the number of question marks and 45 for the number of exclamation marks.

Lastly, the overall tendency for the sentiment score is that the comments are more positive than negative, with a mean of .065 (*SD*=.26).

Table 6: Summary statistics of the markers found in literature detected in the FOK!forum data.

	word count	**pausality**	**lexical diversity**	**# of question marks**	**# of exclamation marks**	**sentiment**
Min.	0	0	0	0	0	-1
1st Qu.	7	1	.85	0	0	0
Median	15	1	.96	0	0	0
Mean	31.71	1.48	.91	.34	.13	.065
3rd Qu.	33	2	1	0	0	.18
Max.	6468	69	1	77	45	1
Sd.	78.29	1.40	.12	.82	.59	.26

Spearman's rank order correlation In order to assess the monotonic relationship between the continuous markers, Spearman's rank order correlation was computed. Table 7 represents the Spearman's correlation coefficient. All coefficients are statistically significant, meaning that the null hypothesis of no relationship between the markers can be rejected. Most markers have a slight correlation, with a correlation coefficient r between -.2 and .2 ($p < .001$) (Burns & Burns, 2008). The relationship is so low as to be random. Pausality and lexical diversity have a low correlation, $r(113524) = .37$, $p < .001$. This means that they have a weak relationship. Word count and pausality have a moderate correlation, $r(113524) = .47$, $p < .001$. The only markers with a high correlation, and thus a substantial relationship, are lexical diversity and word count, $r(113524) = .82$, $p < .001$.

Table 7: Spearman's rank order correlation coefficient r for all continuous markers.

	1	2	3	4	5	6
1. word count	-					
2. pausality	.469*	-				
3. lexical diversity	-.818*	-.367*	-			
4. # of question marks	.140*	.112*	-.130*	-		
5. # of exclamation marks	.099*	.128*	-.094*	.026*	-	
6. sentiment	.143*	.063*	-.108*	-.011*	.057*	-

* p < *.001*

5.2.2 Categorical variable

Descriptive statistics The percentage of comments per class ('Yes' or 'No') of the categorical marker personal perspective is visible in Figure 5. This graph shows that there are more comments with no personal perspective (N=66487) than comments that do contain a personal perspective (N=47039).

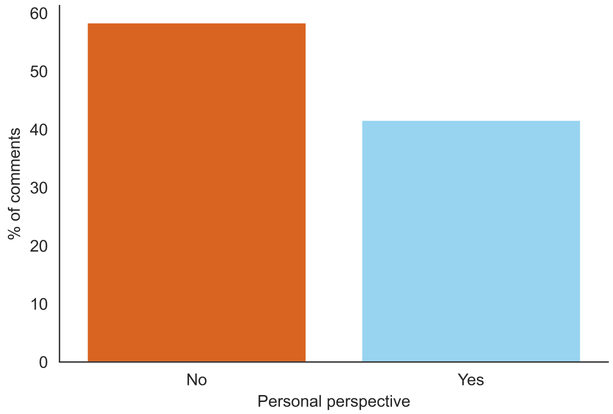
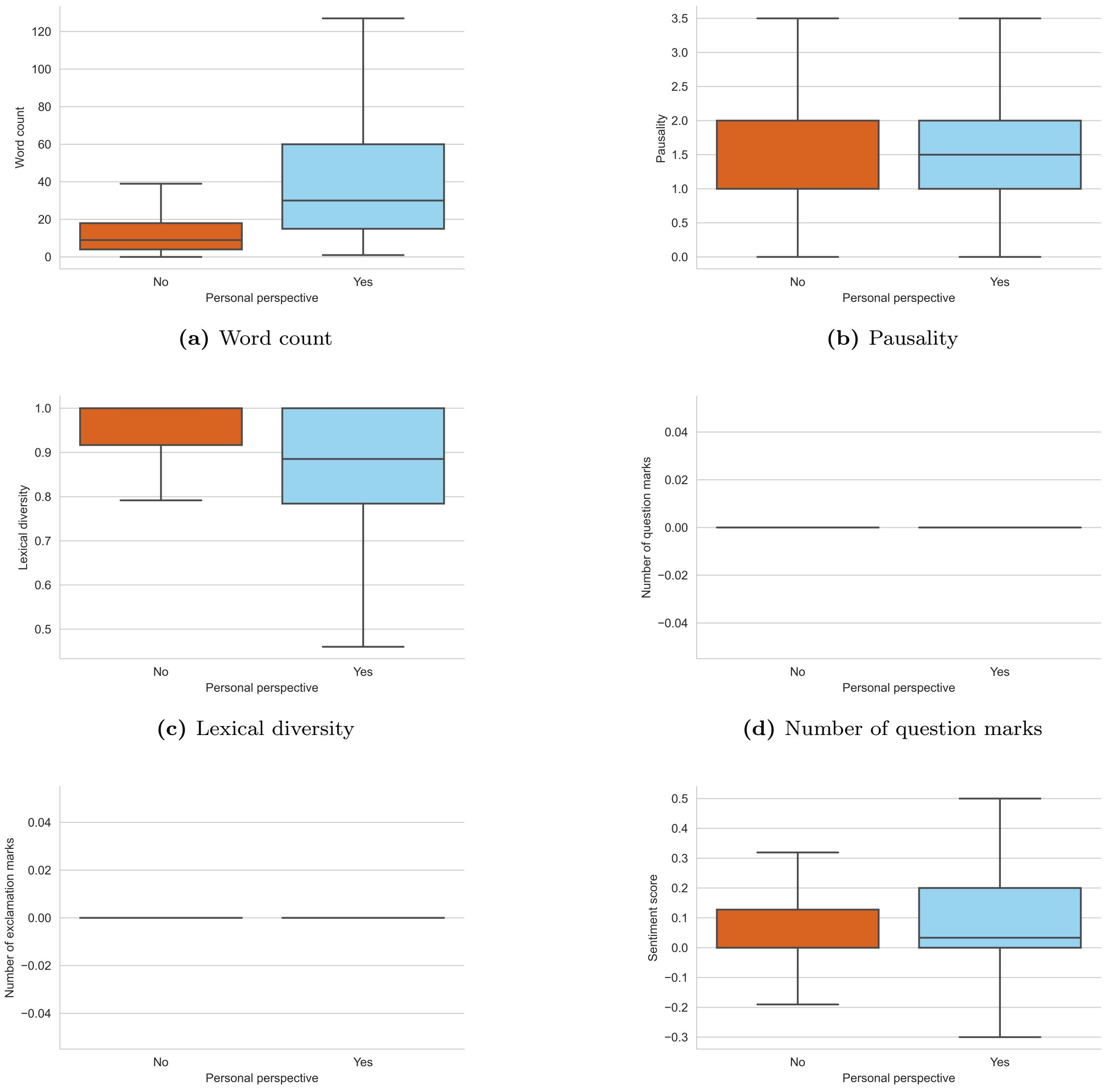


Figure 5: Percentage of comments per category of marker the automatically coded marker personal perspective.

Inter-rater reliability Cohen's kappa κ was computed to assess the agreement between the manually coded marker `V3` personal perspective and the automatically coded marker personal perspective. According to the guideline for interpreting kappa from Landis and Koch (1977), there was an almost perfect agreement between the computer and the manual coder, κ = .87 (95% CI, .85 to .88), $p <$ .001. To put this in more concrete numbers, 246 of the 4063 rows that were compared were not in agreement.

Comparison A box plot without displaying outliers was computed for all markers per class of the marker personal perspective to investigate the relationship between a comment containing a personal



(a) Word count

(b) Pausality

(c) Lexical diversity

(d) Number of question marks

(e) Number of exclamation marks

(f) Sentiment score

Figure 6: Box plots without outliers of all markers compared per class of the marker ‘personal perspective’.

perspective and the continuous variables. Figure 6a shows that the median of the marker word count, indicated by the middle line in the box plot, is higher for comments that do contain a personal perspective. The difference in word count is 21. The box and the whiskers for the positive ('Yes') class are also longer, indicating that the word count is more dispersed.

The boxes and their whiskers from the box plots for pausality (see Figure 6b) are similar, meaning that the dispersion of the pausality per class is similar. The difference is that the median pausality in the negative ('No') class is .50 lower than the median of the positive class.

The median lexical diversity is higher for the negative class, with a value of 1.0, compared to the positive class. Besides, the box and the lower whisker are longer, indicating more dispersed lexical diversity within the positive class. However, most comments in both classes still have relatively high lexical diversity.

The box plots for both the number of question marks (see Figure 6d) and the number of exclamation marks (see Figure 6e) have a median of 0 in both classes, meaning that most comments do not contain any question or exclamation marks. The fact that the box and its whiskers are all on the same line means that all comments, except for the outliers that are not shown here, contain 0 question and exclamation marks, and being in the positive or negative class does not make a difference.

Lastly, the difference between classes for the sentiment score is also not very big (see Figure 6f), with a median of 0 for the negative and a median of .03 for the positive class, respectively. The box and its whiskers of the positive class are also longer, meaning that the data is more scattered than in the negative class. The positive class also has a higher maximum and lower minimum sentiment score than the negative class.

6 Discussion and conclusion

6.1 Interpretation of results

This research aimed to create an overview of the markers that were proven to help detect credible information by existing literature, and to examine whether and how these markers could be applied to the automatic credibility evaluation of user-generated sexual health information. The complete overview of all markers can be found in Chapter 2.2. This section will discuss the interpretation of the results of automating those markers and investigating their relationship.

6.1.1 Manually coded marker `V3` personal perspective

First, the manually coded marker `V3` personal perspective was automated using logistic regression and an SVM. Looking at the accuracy (see Table 5), both models performed just as well in predicting personal perspective. The F1-scores, which consider the data distribution, indicated that the logistic regression model performed just a little better. However, the difference was only .01. In general, the high results of both models indicate their generalizability, and using one model over the other would not make a big difference. However, looking at the speed and interpretability of both models, it would be better to use logistic regression. The speed refers to "the computational costs involved in generating and using the given classifier" and interpretability is "the level of understanding and insight that is provided by the classifier' (Han et al., 2012). In this case, the logistic regression classifier was circa 58 times quicker to train (see Table 5) and the coefficients that the logistic regression model provided (see Figure 4) made the model better interpretable.

These coefficients also indicate that the same words, namely ‘ik’, ‘me’, ‘mij’, ‘mijn’, ‘m’n’, and ‘mn’, were relevant for classifying the manual marker `V3` personal perspective marker as the automatic personal perspective marker. The almost perfect inter-rater reliability of the manually coded comments and the automatically coded comments confirm that modelling personal perspective by simply finding the comments with the words mentioned above would just as well be a good strategy. The main reason why the inter-rater reliability was not perfect, was that some comments were written from a personal perspective, but it did not contain any of the indicative words like ‘ik’ (e.g., “had toch gevraagd om een vrouwen topic en niet om een topic waar mannen over hun eigen penis gaan praten?”).

In summary, the results show that machine learning models are helpful in modelling the manually coded marker `V3` personal perspective on unseen data. This indicates that it could be worthwhile to use similar methods to model other manually coded markers – for example, the other markers from the dataset discussed in Chapter 3.2 – on unseen data. This would save much time compared to manually coding markers, especially since the FOK! forum dataset is enormous. However, the inter-rater reliability shows that more simple methods can also be an accurate way of modelling markers on data. When considering these two methods (machine learning versus conventional methods), the choice thus depends on the type of marker and the presence of a labelled dataset.

6.1.2 Automatic markers

The markers that were automatically modelled on the data were word count, lexical diversity, pausality, number of question marks, number of exclamation marks, and sentiment. For these markers, Spearman’s rank order correlation was determined. The only markers with a strong negative correlation are word count and lexical diversity, meaning that if one marker increases, the other decreases. This is in line with the literature (see Table 2). Almost all studies that investigated the effect of word count on credibility – except for the study of Alrubaian et al. (2016) that concluded word count did not affect credibility – mentioned that a higher word count indicated a lower credibility (Addawood et al., 2019; Janze & Risius, 2017; Zhao et al., 2021; Zhou et al., 2004, 2003). Futhermore, all studies that mentioned the effect of lexical diversity on credibility stated that a higher lexical diversity lead to a higher credibility (Zhou et al., 2004, 2003).

Additionally, the only two markers with a moderate positive correlation were word count and pausality, meaning that as the word count increased, the pausality also increased. This only aligns with the study of Addawood et al. (2019), who found that a high word count and high pausality cause low credibility. However, the studies of Zhou et al. (2003) and Zhou et al. (2004) showed the opposite, namely that high pausality indicated higher credibility.

Furthermore, pausality and lexical diversity have a weak relationship in the negative direction. This means that a high pausality and a low lexical diversity should lead to low credibility. The literature does not support this because multiple studies found that both low pausality and low lexical diversity lead to low credibility (Zhou et al., 2004, 2003). Lastly, all other continuous markers only showed a slight correlation, meaning that a relationship is so small as to be random (Burns & Burns, 2008).

The markers were also compared to the categorical marker personal perspective using box plots (see Figure 6). The most noteworthy difference between the two classes (‘Yes’ or ‘No’) is visible within the marker word count. The positive class has an overall higher word count than the negative class, which could indicate that a higher word count also increases the chances of a personal perspective. This is not supported by literature, as the studies of both Addawood et al. (2019) and Zhou et al. (2004) show that using self-reference increases credibility. That would mean that a high word count would also lead to higher credibility, but this is not the case, according to the literature. It is, however, not possible to say with certainty if the word count is a predictor of personal perspective since the data did not meet the assumptions of various statistical tests. However, it is something worth investigating in more depth in future research. Additionally, all other markers did not show a noticeable difference between the two classes.

Considering all the results, it seems that only the relation between word count and lexical diversity and word count and pausality are (somewhat) in line with the results from other studies on automatic credibility assessment. The fact that the other markers, number of question marks, number of exclamation marks, sentiment, and personal perspective, did not show meaningful relationships may indicate that these markers are not generalizable to sexual health data. However, due to the limitations of this study, further research is needed to confirm this.

6.2 Further limitations and future work

Several limitations and weaknesses of this study need to be addressed, and are followed by ideas for future research.

First, due to a lack of a labelled dataset, it is difficult to confirm if and how the markers help detect credible sexual health information. Nonetheless, this study gave a good insight into how these markers influence each other and if this was in line with current research. Since most markers do not have a strong correlation, this could indicate that only using linguistic markers to assess credibility is not enough. This is also in line with the results of several studies mentioned in 2. For example, Castillo et al. (2011) stated that "text-based features are not enough by themselves for [assessing credibility]." Zhang and Ghorbani (2020) even went as far as distinguishing between nine different types of features for fake news representation and identification, which they divided into three overarching categories: creator/user-based features (user profiling features, user credibility features, behaviour-based features), news/content-based features (linguistic and syntactic features, style based features, visual-based features), and social context-based features (network-based features, impact-based features, temporal-based features). Investigating all these different types of features was beyond the scope of this research. Besides, this study was constrained by the dataset, which did not contain any information on, for example, users. Therefore, future research and a more comprehensive dataset are needed to establish whether the linguistic markers, in combination with other markers, would be able to aid in the detection of credible sexual health information.

There are several examples of markers that could be investigated with a more comprehensive dataset. Zhao et al. (2021) found that authors of misinformation created significantly more threads and replies than authors of legitimate information, implying that users creating misinformation tended to be more engaged in the discussions in the community. Zhang and Ghorbani (2020) also propose looking at the number of threads and replies a user created to reflect users' activity. Besides, several articles, such as Addawood et al. (2019) and Agichtein et al. (2008), mentioned that the fraction of posts with, for example, URLs and question marks per user would be a better indicator than simply counting the total number of posts containing one of these elements.

Besides, it is evident from the overview of the markers in Chapter 2 that the results from different studies do not always agree. One reason for this could be that the markers in these studies are often applied to different types of online media. For example, a lot of studies applied the markers to tweets (e.g., Addawood et al., 2019; Olteanu et al., 2013; Zubiaga et al., 2017). Tweets only contain a limited number of characters, so the results may be different from those for websites where both short and longer posts are allowed. In future research, the markers could be applied to different media types. With a labelled dataset for each of these media types, the markers could be compared to see if they impact credibility differently for different media types.

Second, there is an ethical implication of only using content-based markers. Due to a difference in the educational level or Dutch not being a person's first language, there is a risk that an automated credibility detection system is biased towards people from different backgrounds or with different educational levels. The system sees these people as less credible solely based on their language and writing skills. Once again, this underpins the idea that textual markers should be combined with other types of markers, for which a more comprehensive dataset is needed. This is, however, not to say that the textual markers are also an essential part of credibility detection, which has become evident from all the research (see 2 that has already been done on these markers.

Third, the markers found in the literature significantly impacted credibility detection in predominantly English texts. For this reason, one could argue that these markers do not apply to Dutch texts. However, English and Dutch are both West-Germanic languages that share many similarities, both lexically and grammatically (Harbert, 2007). It is beyond the scope of this study to take the differences between the languages into account. Further research is needed to establish the impact of these differences on the results.

This also ties in with the fact that the methodological choices were constrained by the accuracy of the tokenizer `spaCy`. Although the `spaCy` model can process English texts very accurately, it is not as accurate for Dutch texts (spaCy, n.d.-a, n.d.-b). For this reason, the decision was made to not automate some markers, like content words, since the `spaCy` module was not equipped enough to detect the right content word for each token.

Lastly, an idea for future research within the scope of the dataset is to investigate these markers per thread. For example, Janze and Risius (2017) found that the average number of comments on a post containing a question mark decreases the possibility of that post containing fake news. Besides, Agichtein et al. (2008) found that bad questions attract more bad answers, whereas good answers are more likely to be written in response to good questions. They also found that the ratio between the question and the answer length was significant in determining the quality of answers. These examples make a good case to research complete threads in the future, instead of every post or comment separately.

6.3 Conclusion

In summary, this research aimed to create an overview of the content-based markers that were important for detecting credible information according to existing literature, and to examine whether and how these markers could be applied to the automatic credibility evaluation of user-generated sexual health information.

The results from this research show that some markers can be modelled on the data in various ways, using machine learning approaches or more conventional methods. Besides, to the author's knowledge, this was the first study that succeeded in making such a comprehensive overview of content-based markers.

Although this research is by no means a comprehensive work of all possible credibility markers, it is a start in understanding if the automatic credibility evaluation of sexual health information is possible. The lack of relationships between most markers may indicate that they are not applicable to the automatic credibility assessment of sexual health information. In contrast, the relations between the markers word count and lexical diversity and word count and pausality do show agreement with existing literature to a great extent, which is promising. Therefore, it can be concluded that further research is needed to determine the direct effects of the markers on the credibility of sexual health information.

Altogether, the results of this study led to an enormous amount of ideas for future research, which should be used as a source of inspiration and as food for thought for anyone wishing to develop an effective and comprehensive credibility assessment framework.

References

Addawood, A., Badawy, A., Lerman, K., & Ferrara, E. (2019). Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 15–25).

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 183–194).

Alrubaian, M., Al-Qurishi, M., Hassan, M. M., & Alamri, A. (2016). A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 661–674.

Artusi, R., Verderio, P., & Marubini, E. (2002). Bravais-pearson and spearman correlation coefficients: meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, *17*(2), 148–151.

Astivia, O. L. O., & Zumbo, B. D. (2017). Population models and simulation methods: The case of the spearman rank correlation. *British journal of mathematical & statistical psychology*, *70*(3), 347-367.

Attwood, F., Barker, M. J., Boynton, P., & Hancock, J. (2015). Sense about sex: Media, sex advice, education and learning. *Sex Education*, *15*(5), 528–539.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55.

Burns, R. B., & Burns, R. A. (2008). *Business research methods and statistics using spss.* London, England: SAGE Publications.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684).

Chao, C. C. (2017). Correlation, point-biserial. In M. Allen (Ed.), *The sage encyclopedia of communication research methods.* SAGE Publications.

Chou, W. S., Prestin, A., Lyons, C., & Wen, K. (2013). Web 2.0 for health promotion: reviewing the current evidence. *American journal of public health*, *103*(1), 9–18.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). *Introduction to information retrieval.* Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

De Graaf, H., Van den Borne, M., Nikkelen, S., Twisk, D., & Meijer, S. (2017). *Seks onder je 25e: Seksuele gezondheid in nederland anno 2017.* Delft: Eburon Uitgeverij.

Doornwaard, S. M., den Boer, F., Vanwesenbeeck, I., van Nijnatten, C. H., Ter Bogt, T. F., & van den Eijnden, R. J. (2017). Dutch adolescents' motives, perceptions, and reflections toward sex-related internet use: Results of a web-based focus-group study. *The Journal of Sex Research*, *54*(8), 1038–1050.

Eysenbach, G. (2008). Credibility of health information and digital media: New perspectives and implications for youth. In M. Metzger & A. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 123–154). Cambridge, MA: The MIT Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378-382.

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 80-87). New York, NY: Association for Computing Machinery.

Han, J., Pei, J., & Kamber, M. (2012). *Data mining: concepts and techniques* (3rd ed.). Elsevier.

Harbert, W. (2007). *The germanic languages.* Cambridge, England: CUP.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer science review*, *41*, 1135–1140.

Janze, C., & Risius, M. (2017). Automatic detection of fake news on social media platforms. In *Pacis 2017 proceedings.*

Jiaying, L., Song, S., & Zhang, Y. (2021). Linguistic features and consumer credibility judgment of online health information.

Kanuga, M., & Rosenfeld, W. D. (2004). Adolescent sexuality and the internet: the good, the bad, and the url. *Journal of Pediatric and Adolescent Gynecology*, *17*(2), 117–124.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4).

Lai, J. (2022). *The extent of sentiment in sexual health information between moderated and non-moderated websites.* Unpublished master's thesis, Utrecht University, Utrecht.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Merriam-Webster. (n.d.). Credible. In *Merriam-webster dictionary.* Retrieved July 19, 2022, from `https://www.merriam-webster.com/dictionary/credible`

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 441–450).

Mukherjee, S., Weikum, G., & Danescu-Niculescu-Mizil, C. (2014). People on drugs: Credibility of user statements in health communities. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 65–74).

Nasteski, V. (2017, 12). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51-62.

Nikkelen, S. W. C., van Oosten, J. M. F., & van den Borne, M. M. J. J. (2020). Sexuality education in the digital era: Intrinsic and extrinsic predictors of online sexual information seeking among youth. *The Journal of Sex Research*, *57*(2), 189-199.

Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. (2013). Web credibility: Features exploration and credibility prediction. In P. Serdyukov et al. (Eds.), *Advances in information retrieval* (pp. 557–568). Berlin, Heidelberg: Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Blondel, M.

(2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Schwarz, J., & Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the international conference on human factors in computing systems* (p. 1245–1254). New York, NY: Association for Computing Machinery.

scikit-learn. (n.d.-a). *User guide: Support vector machines.* Retrieved July 18, 2022, from `https://scikit-learn.org/stable/modules/svm.html#`

scikit-learn. (n.d.-b). *User guide: Text feature extraction.* Retrieved July 15, 2022, from `https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction`

scikit-learn. (n.d.-c). *User guide: Tuning the hyper-parameters of an estimator.* Retrieved July 18, 2022, from `https://scikit-learn.org/stable/modules/grid_search.html#`

Shami, M., & Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech communication*, *49*(3), 201–212.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *Database system concepts* (Vol. 5). New York, NY: McGraw-Hill.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (indiacom)* (p. 1310-1315).

spaCy. (n.d.-a). *Dutch.* Retrieved June 27, 2022, from `https://spacy.io/models/nl`

spaCy. (n.d.-b). *English.* Retrieved June 27, 2022, from `https://spacy.io/models/nl`

Spark Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. *Journal of Documentation*, *28*(1), 11–21.

Stommel, M., & Dontje, K. J. (2014). *Statistics for advanced practice nurses and health professionals.* New York, NY: Springer Publishing Company.

Wawer, A., Nielek, R., & Wierzbicki, A. (2014). Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd international conference on world wide web* (p. 1135–1140). New York, NY: Association for Computing Machinery.

Wikipedia. (n.d.). *FOK!* Retrieved June 02, 2022, from `https://nl.wikipedia.org/wiki/FOK!`

Wise, K., Hamman, B., & Thorson, K. (2006). Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication*, *12*(1), 24–41.

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, *28*(14), 3866-3878.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 1-26.

Zhao, Y., Da, J., & Yan, J. (2021). Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, *58*(1), 1-24.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, *13*(1), 81–106.

Zhou, L., Twitchell, D., Qin, T., Burgoon, J., & Nunamaker, J. (2003). An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th* *annual hawaii international conference on system sciences, 2003* (pp. 1–10).

Zubiaga, A., Liakata, M., & Procter, R. (2017). Exploiting context for rumour detection in social media. In *International conference on social informatics* (pp. 109–123).