



Universiteit
Leiden

The Netherlands

Bachelor Informatica & Economie

Analyzing narratives about
cancer screening on social media

Emma Michielsens

Supervisors:

Suzan Verberne & Marco Spruit

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)

www.liacs.leidenuniv.nl

19/12/2024

Abstract

For this research, I looked at how automatic text analytics methods can help to identify and categorise narratives about cancer screening on social media. I used three supervised classifiers to categorize 4885 tweets into subcategories for inclusion, cancer type and topic. I also used one unsupervised classifier to perform sentiment analysis on these 4885 tweets. I found that supervised classifiers can effectively categorise tweets and sentiment analysis can provide further insights. For cancer type, the key descriptors for the general or unclear subcategory are ‘bevolkingsonderzoek’ (population screening) and ‘kankerscreening’ (cancer screening). For the other cancer types, the key descriptors are the name of the cancer type and the associated screening test. The key descriptors for the topic category are ‘ik’ (I) and ‘uitslag’ (result) for the personal subcategory and ‘darmkanker’ (colorectal cancer) and ‘screening’ (screening) for the non-personal subcategory. Sentiment analysis revealed that, in general, cancer screening is discussed more negatively than positively on social media. These findings can give insight into the cancer screening narratives and help organize strategies on how to inform Dutch citizens to participate in national cancer screening programmes.

Contents

1	Introduction	1
2	Related work	2
3	Data	3
3.1	Data collection	3
3.2	Data labelling	3
3.3	The datasets	6
4	Methods	7
4.1	Machine learning models	7
4.2	Key features	9
4.3	Analysis of the large dataset	9
5	Experiments and results	10
5.1	Results	10
5.2	Key features	16
5.3	Analysis of the large dataset	17
6	Discussion	20
7	Conclusions and further research	21
	References	24

1 Introduction

Every Dutch citizen is invited to participate in one of the national cancer screening programmes at a specific stage in their lives. The breast cancer screening is for women aged between 50 and 75, who are invited every two years for a mammogram. The cervical cancer screening is for women between the ages of 30 and 60 years, who are invited every five years for a smear test, or they can request a self-sampling device. The colorectal cancer screening is for men and women between 55 and 75 years old, who get sent a home kit to collect a stool sample.¹ These screenings are offered free of charge by the government. However, not everyone who receives this invite participates. The IKNL (Integraal Kankercentrum Nederland) keeps track of the performance of the three national cancer screening programmes. In 2021 72.5% of the invited citizens participated in the breast cancer screening, 54.8% participated in the cervical cancer screening and 70.6% in the colorectal cancer screening [Ned22, Ned24].²

A potential reason for these low participation rates could be found on social media. Social media analysis can give valuable insight into whether people generally think positively or negatively about cancer screening. It can show if people talk about personally experiencing discomfort with cancer screening procedures or spread negative opinions based on news reports. It can also show if people either encourage participation or spread hesitation and fear towards cancer screening. Analysing how people talk about cancer screening on social media can help to address misconceptions and develop targeted strategies to promote participation. Automatic text analytics methods could offer a way to quickly uncover these insights and better understand how cancer screening is discussed on social media.

To guide this research, my main research question is:

How can we use automatic text analytics methods to identify and categorise narratives about cancer screening?

To answer this question I address three sub-questions:

1. How well can automatic text analytics methods classify social media posts about cancer screening?
2. Which words serve as key descriptors for categorising these posts?
3. What is the distribution of these categories within a large dataset of 4885 social media posts?

I will use automatic text analytics methods to classify tweets about cancer screening. To evaluate the effectiveness of the models, I will look at the precision and recall scores of the test data after the models are trained. I will also look at the top key features for multiple categories to find which words serve as key descriptors for these categories. I will show the distribution of these categories within the original dataset and the large dataset of 4885 social media posts. The code I developed for this research is available on GitHub.³

¹<https://www.rivm.nl/en/population-screening-programmes>

²<https://iknl.nl/en/screening>

³<https://github.com/Emmamich/Bachelor-Thesis-Emma-Michielsens>

2 Related work

There is a large amount of research on analysing Twitter data within the health domain. There have been studies on how Twitter data can be used for public health research in the context of surveillance, detection and prediction of public health trends and conditions [EODLILE20] and also how sentiment analysis can be used to assess public health concerns [JCWG15].

Regarding cancer, multiple studies have explored various aspects of how this is discussed on social media. Koval et al. (2017) explored this by applying text mining techniques to gain insights from a large dataset of tweets. The researchers conducted a study on Twitter discussions related to cancer. The results from their clustering process identified prominent topics within cancer-related tweets and quantified the distribution of tweets across these topics for six specific cancer types. Interestingly, the top six cancers discussed on Twitter also aligned with the most common cancers in 2016, with the exception of ovarian cancer. They found that this difference may be due to ovarian cancer being the fifth leading cause of cancer-related death among women. The analysis further showed that breast cancer was the most frequently discussed cancer type, accounting for 58.6% of the tweets in the top six categories. The study also revealed significant spikes in tweet activity related to specific cancers that corresponded with annual cancer awareness months for those types. Another insight was that celebrities play a substantial role in shaping the frequency and content of cancer-related conversations on social media, highlighting their potential in raising public awareness around cancer prevention and treatment [KLL17].

Another study also found that breast cancer was the most commonly discussed cancer on Twitter along with lung, prostate and colorectal and it found that most tweets were about patients' experience with treatment [CCJ+16]. For colorectal cancer, one study found that most tweets were about news articles and risk or prevention [POP+16]. For cervical cancer, another study found that during cervical cancer awareness month, professional health organizations were responsible for 20.7% of the tweets and just 11.2% of all the tweets in their sample featured personal stories from cervical cancer patients [TSV+18]. One study has found that for breast cancer, the largest portion of news stories were classified as 'real-life story' with 52.5% and that 5.08% of the total stories mention prevention and 19.7% mention early detection/screening. They also found that news stories classified as 'rumours' were shared 3.29 times more than those scientifically correct [BMC21]. There has been another study to detect rumours on Twitter within the health domain which found a system to correctly identify around 90% of the rumours [SGP+18]. Another study has found that from their sample of tweets about gynecologic cancer, approximately 30% contained misinformation. In addition, they found that tweets about cancer treatment were found to contain a higher proportion of misinformation compared to those about prevention [CWP18]. A study in Japan found that within their dataset of tweets containing the Japanese word for cancer, 44% contained misinformation [KTK+23].

Sentiment analysis has also been applied to social media discussions about cancer. One study found that tweets related to colonoscopies, a colorectal cancer screening test, were more likely to have negative sentiment than positive sentiment. On the contrary, tweets related to mammograms, a breast cancer screening test, were more likely to have a positive sentiment. The distribution of positive and negative sentiment for pap smears, a cervical cancer screening test, were not significantly different. For all three tests, at least 75% was classified with a neutral sentiment [MBLS17].

Another study found that Twitter can serve as a supportive platform for breast cancer patients

and that positive experiences were shared in regard to patient treatment [CJJ+18]. For cervical cancer, one study found that positive tweets, focussing on the benefits of cancer screening, increased promotion and retweets [LRL+19]. For colorectal cancer, a study used sentiment analysis to find which screening method people have a more positive view on [CLC23].

Within the Netherlands, there has been research on the different perspectives on cancer screening, which found that some respondents prefer not knowing about potential cancer without symptoms, have relatively low expectations of screening lowering cancer death risk and some respondents brought up the potential for receiving test results that could be either falsely positive or falsely negative. This research used online questionnaires and interviews in the city of The Hague to gather their data [BBC+22]. News media and social media content could also be used to gather data on the opinions of Dutch citizens about cancer screening. SENTENCES is a project for social media analysis in the Netherlands to promote cancer screening [HVdBD23].⁴ That study has not been published yet but the data is shared for this project. The results of this study can help organize strategies on how to inform Dutch citizens about participating in population-based screening for cancer.

3 Data

3.1 Data collection

A total of 4885 Dutch tweets were gathered by the project SENTENCES, containing the keywords ‘bevolkingsonderzoek’ (population screening), ‘kankerscreening’ (cancer screening), ‘uitstrijkje’ (smear test), ‘zelfafnametest’ (self-sampling test), ‘zelfafnameset’ (self-sampling kit), ‘mammografie’ (mammography), ‘mammogram’ (mammogram), ‘ontlastingstest’ (stool test), ‘poepstest’ (poop test), ‘baarmoederhalskanker screening’ (cervical cancer screening), ‘borstkanker screening’ (breast cancer screening) or ‘darmkanker screening’ (colorectal cancer screening). These tweets spanned from January 1, 2010, to October 31, 2022. From these 4885 tweets, a subset of 1629 tweets was selected to label.

3.2 Data labelling

A total of 12 students categorized the tweets as part of the SENTENCES project. These 12 students were divided into four subgroups with 3 students each. Each group had to categorize 407 tweets, about 135 tweets per person. Of these 407 tweets, 44 tweets had to be double-coded, which is 44 tweets per person. In double-coding, all three members of the subgroup coded the same set. They calculated the intercoder reliability with Krippendorff’s alpha. After, they discussed the categorization rules, adjusted them where needed and did another round of double-coding. Then they calculated Krippendorff’s alpha again, receiving sufficient scores: 0.89 for inclusion, 0.95 for cancer type, 0.77 for topic and 0.51 or higher for the other categories.

The students categorised each tweet based on specific criteria. Initially, they checked if the tweet met the requirements: it should be a post in Dutch, including a reference to screening for cervical, breast, or colorectal cancer as part of the population screening programme in the Netherlands. After it was determined the tweet should be included, the tweet was categorised into different

⁴<https://projecten.zonmw.nl/nl/project/sentences-social-media-analysis-promote-cancer-screening>

subcategories. They determined the type of cancer discussed in each tweet and categorised the tweets into six different topics. Additionally, they identified if the tweet included information about the severity of developing one of the three types of cancer, the perceived benefits of cancer screening and the perceived barriers. The students also examined whether the tweet contained an indication to participate or not participate in cancer screening. Lastly, they noted if the tweet referred to the extent to which someone is able to perform cancer screening. They stored their results in a table, where each row has a tweet with columns of categories behind it. In each column, there is a number that stands for one of the subcategories. These results can be seen in the labelled dataset in GitHub.⁵

Below, each of the categories that are relevant for this study, will be discussed and shown how the students were instructed to decide which tweet should get which subcategory. For more information, the code instructions are on GitHub under ‘Codeboek’, which was made by the researchers for the SENTENCES project.

Inclusion For the category inclusion, there are 2 subcategories:

0. Exclusion

1. Inclusion

Posts that clearly refer to cancer screening in Belgium rather than the Netherlands and posts that clearly refer to other population screenings, such as lung cancer should be excluded. Posts identified by "RT" or "@" at the beginning of the tweet, should also be excluded. Posts that discuss alternate screening options, thermography for example, to those offered by the population screening, should be included. Tweets should only be excluded if it is clear that the cancer screening mentioned is not part of the population screening programme. When it is not certain, the tweet should be included. Figure 1 shows the distribution of how the students labelled the data as exclusion or inclusion.

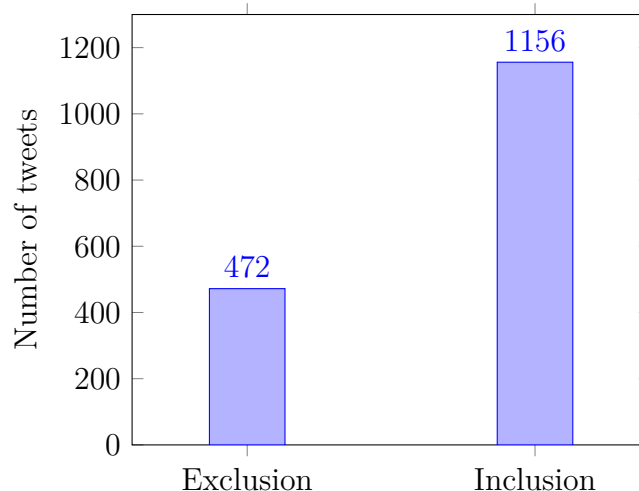


Figure 1: Distribution of the tweets labelled as exclusion or inclusion

⁵<https://github.com/Emmamich/Bachelor-Thesis-Emma-Michielsens>

Cancer type For the category cancer type, there are 4 different subcategories:

0. General or unclear
1. Cervical cancer
2. Breast cancer
3. Colorectal cancer

The tweet is labelled as ‘general or unclear’ when cancer population screening is discussed in general, when it is unclear which cancer type is meant, or when there is a reference to more than one type of cancer. Figure 2 shows the distribution of the labelled data for cancer type.

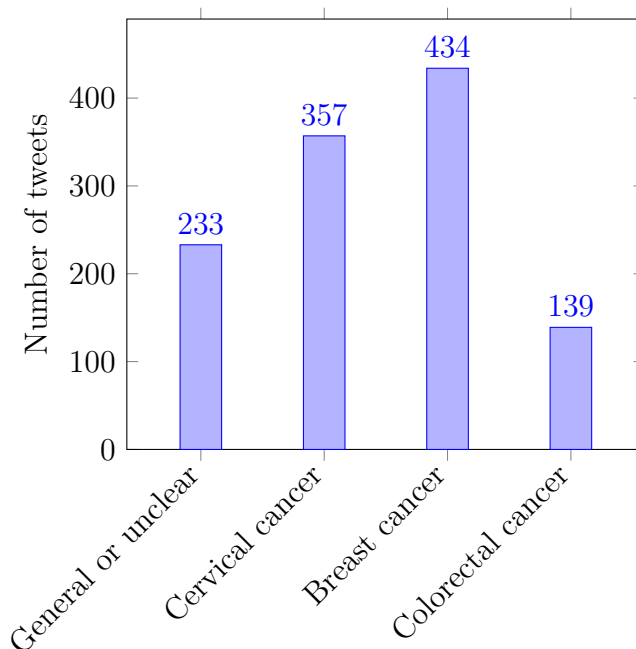


Figure 2: Distribution of the labelled tweets for the category cancer type

Topic For the category topic, there are 6 different subcategories:

0. Other
1. Legislation and regulations
2. Events
3. Science
4. News
5. Personal

The tweet is labelled as ‘Personal’ when it is about a personal situation of the writer or about someone they know. When a tweet is a personal opinion, it should be labelled as which subcategory the opinion is about. Figure 3 shows the distribution of the labelled data for topic. It shows that the ‘personal’ subcategory contains substantially more indices than the other subcategories, which is why I looked at the distribution of topic with all subcategories, except personal, grouped into the global subcategory ‘not personal’, which is shown in Figure 4.

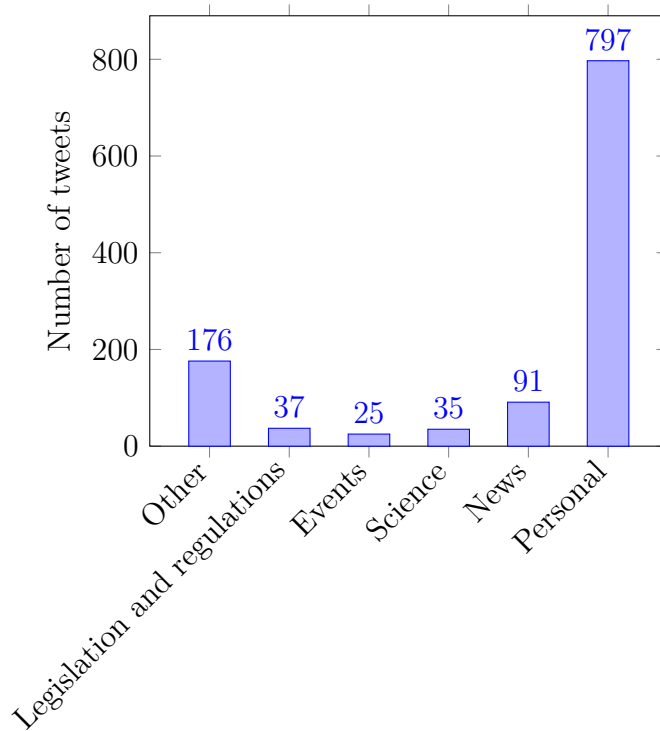


Figure 3: Distribution of the labelled tweets for the category topic

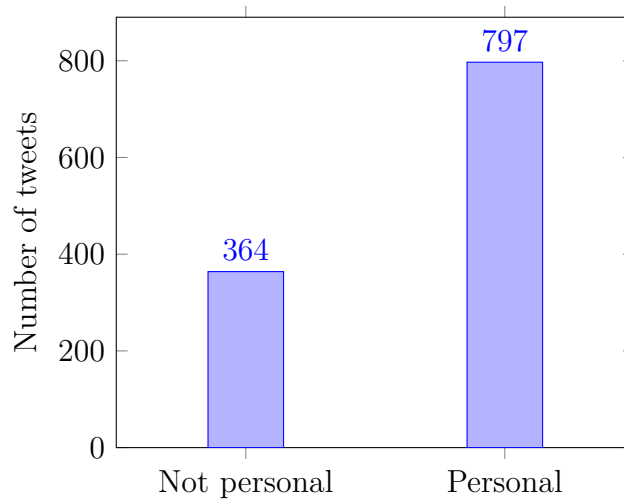


Figure 4: Distribution of the labelled tweets for the category topic with all subcategories except personal grouped into the global subcategory ‘Not personal’

3.3 The datasets

Table 1 shows part of the first four rows of the labelled dataset, with their corresponding row numbers, tweets in the ‘Text’ (text) column and labels for the ‘Inclusie’ (inclusion), ‘Kankersoor’ (cancer type) and ‘Topic’ (topic) columns. As I mentioned in Section 3.1, the tweets in the labelled dataset are a subset of the total tweets that were gathered. Table 2 shows part of the first four

	Text	Inclusie	Kanker-soort	Topic
1	Zo pijnlijk #mammografie niet alleen mijn borst maar ook nek moest er onder. Moest mij niet aanstellen! 3 weken grote blauwe plekken gehad en onderzoek mislukt. Moet toch gebeuren maar weet niet hoe. @bvo_nederland	1	2	5
4	”Na de tandarts (maandag) en de herhaalprik (dinsdag) kwam woensdag een uitnodiging voor het bevolkingsonderzoek borstkanker 😞 Willen jullie op 11 november voor mij duimen ‘s middags? 😊	1	2	5
7	”Het leed dat #Tinder heet: Wanneer je je huisarts belt voor een soatest...En ze ff moet overleggen of je niet toch ook een uitstrijkje nodig hebt. #akward maar zo verstandig.”	0		
10	”Binnenkort krijgen vrouwen van 30+ een setje toegestuurd om zelf een uitstrijkje (baarmoederhalskanker) te maken. Ik probeer me dat voor te stellen maar het lijkt me eerlijk gezegd nog niet zo eenvoudig.”	1	1	1

Table 1: Sample of tweets collection with labels

rows of the large dataset of 4885 tweets, which have not been labelled. I will use both the data of the labelled dataset and the large dataset to generate the results for this study.

4 Methods

4.1 Machine learning models

For each category described in Section 3.2 – inclusion, cancer type and topic – I trained a separate classifier. For sentiment analysis, I used a pre-trained model. So, I employed three supervised models and one unsupervised model. I implemented the analysis in Google Colab and my code can be found on GitHub. ⁶ First, I imported the necessary libraries and modules. Next, I imported the labelled dataset for the supervised models and the unlabelled dataset with the 4885 tweets for the unsupervised model, both from an Excel file. I replaced the empty cells in the ‘Text’ column with NaN values and then removed these rows. Although I considered removing common stop words from the ‘Text’ column, I decided against it to prevent the loss of important contextual information from the tweets, as such words can sometimes convey sentiment or tone.

⁶<https://github.com/Emmamich/Bachelor-Thesis-Emma-Michielsens>

	Text
1	Zo pijnlijk #mammografie niet alleen mijn borst maar ook nek moest er onder. Moest mij niet aanstellen! 3 weken grote blauwe plekken gehad en onderzoek mislukt. Moet toch gebeuren maar weet niet hoe. @bvo_nederland
2	Volgende week moet mijn meneer (om hem ook maar eens zo te noemen) naar het ziekenhuis, foto's van zijn "tiet" bij mannen kan daar ook iets mis mee zijn en net ook nog post bevolkingsonderzoek , afspraak kijkonderzoek darmen. Gatverdamme
3	Ik heb twee uitnodigingen gekregen voor het bevolkingsonderzoek borstkanker. Moet ik eerst met de ene, dan met de andere..?
4	"Na de tandarts (maandag) en de herhaalprik (dinsdag) kwam woensdag een uitnodiging voor het bevolkingsonderzoek borstkanker 😞 Willen jullie op 11 november voor mij duimen 's middags? 😊 "

Table 2: Sample of large tweet collection without labels

Supervised models For the first binary classifier, I defined the target variable (y) as the ‘Inclusie’ (Inclusion) column. For the second classifier, I defined y as ‘Kankersoort’ (Cancer type). For the third classifier, I defined y as ‘Topic’ (Topic). However, because this model did not perform very well, which I discuss in Section 5.1, I made it a binary classifier by replacing all the topic values of 5 (personal) with 1 and replacing all the other topic values with 0 (non-personal).

For the feature matrix (X) I transformed the text in the ‘Text’ column into numerical values using a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer from sklearn.⁷ This process quantified the words based on their importance within a tweet relative to all tweets. I looked at the unique number of words which was 5595. Because this is a low number of unique words and because it is text data, which means it will have a long tail distribution, I chose for the inclusion classifier to limit the vectorizer to include only the 1000 most important features (words), based on the TF-IDF score, by setting a maximum feature limit. For the cancer type classifier and the binary topic classifier, I set this limit at 1500 to allow for a more detailed and specific analysis, as these classifications require a larger number of features to effectively capture the subtle differences between the subcategories.

For every classification task, I split the dataset into training and testing data, with 80% for training and 20% for testing. I tried both a Random Forest and a Logistic Regression classifier from sklearn, but the latter showed lower precision and recall scores, making it less suitable than the Random Forest classifier.^{8,9} Therefore, I chose the Random Forest classifier for the results presented in Section 5.1.

For the Random Forest classifier, I configured it to use 100 decision trees. I also set a random seed (random state of 42) to ensure that each model produces the same results each time. I then trained the models on the training data. After training, I assessed the models’ performance by predicting outcomes on the test data and created three classification reports to evaluate the performance of each model, which I discuss in Section 5.1. After evaluating the model, I saved it.

⁷https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

⁸<https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

⁹https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LinearRegression.html

Unsupervised model For sentiment analysis, I used a pre-trained model from Hugging Face called `twitter-xlm-roberta-base-sentiment`. This model can classify text into three categories: positive, negative, and neutral and can give a certainty score. The model is based on the XLM-roBERTa-base architecture and was trained on around 198 million tweets, across multiple languages, specifically fine-tuned for sentiment analysis. This fine-tuning was done on eight different languages, not including Dutch, but the model can still effectively handle more languages than these eight.¹⁰ I also tried a pre-trained model from Hugging Face that was trained on Dutch text, but it was trained on book reviews and not on tweets.¹¹ It can also classify the tweets into a positive and negative sentiment but not neutral. I looked at the results of this model on the dataset but found that the other model performed better and all sentiment analyses were therefore performed with `twitter-xlm-roberta-base-sentiment`. Because there is no labelled sentiment data there is no test data and I cannot measure it but I show this in Section 5.1 where I shortly discuss the results of both models.

To perform the sentiment analysis, I first installed the `transformers` library from Hugging Face to get access to the pre-trained models. Then, I used the `transformers` library to create a sentiment analysis pipeline with the `twitter-xlm-roberta-base-sentiment` model and tokenizer. Once this was set up, I applied the sentiment analysis model to the dataset. Looping through each tweet in the 'Text' column, the model predicted for each tweet the sentiment label and associated certainty score. In a new DataFrame, including the original 'Text' column, with a new 'Sentiment' column and 'Certainty Score' column, I stored these results.

4.2 Key features

For the results of Section 5.2, I followed the same steps described in Section 4.1 but replaced the Random Forest classifier with a Linear Regression classifier. A Logistic Regression classifier is more effective for interpreting the results, which is why I used this classifier for Section 5.2, where I analysed the most important features for each class regarding cancer type and topic. Section 5.1 shows that the precision and recall scores of the Logistic Regression classifier are sufficient for the Logistic Regression classifier to be used for analysing the key features. After training the data, I first retrieved all the feature names that were processed by the TF-IDF vectorizer and then all the features' importance scores. I sorted the 50 most important features by their importance score and looped through these features to print the words with their corresponding feature importance scores.

4.3 Analysis of the large dataset

In a new Google Colab notebook, I again imported the necessary libraries and modules, as well as the trained model. I then imported an Excel file containing the large dataset of 4885 tweets without classifications. Following the same preprocessing steps, I removed rows with empty cells in the 'Text' column and used a TF-IDF vectorizer on the 'Text' column. I used the trained model to classify the tweets and added a new column to the dataset to store these predictions. I repeated this step with the other two supervised models to obtain a new dataset with the predictions of all

¹⁰<https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment>

¹¹<https://huggingface.co/DTAI-KULeuven/robbert-v2-dutch-sentiment>

three categories, inclusion, cancer type and topic. I also counted the instances of each subcategory to show in Section 5.3.

5 Experiments and results

To generate the results, I used three supervised classifiers for inclusion, cancer type and topic and one unsupervised classifier for sentiment analysis.

5.1 Results

Inclusion The performance of the inclusion model is shown in Table 3. It demonstrates a clear prioritization of the inclusion class. The recall for this class is 0.99, meaning that the model successfully identifies almost all the true inclusion cases. The precision is 0.72, which means that 28% of the predicted inclusion cases are actually exclusions. The low recall of 0.16 for the exclusion class also shows that the model misses a significant portion of true exclusions. This is acceptable for this study as it was indicated that, in case of doubt, instances should be classified as inclusion rather than exclusion. Therefore, the model’s tendency to prioritise the inclusion class is desirable for this study.

In addition to this, the distribution of the original dataset is also likely to have contributed to the model’s performance. The dataset contained almost 2.5 times more instances of inclusion than exclusion. Because of this smaller number of exclusion cases, the model had fewer opportunities to learn the nuances of identifying exclusions correctly which resulted in the lower recall for exclusions.

	Precision	Recall	Test instances
Inclusion	0.72	0.99	224
Exclusion	0.84	0.16	102

Table 3: Classification report for the supervised model for inclusion and exclusion using a Random Forest classifier

Table 4 shows the performance of the Logistic Regression classifier, but the lower recall score for inclusion and the lower precision score for exclusion indicates that this classifier incorrectly labels inclusion cases as exclusion. As a result, valuable information may be lost, making the Random Forest classifier a better choice for this category in the context of this research.

	Precision	Recall	Test instances
Inclusion	0.73	0.95	224
Exclusion	0.67	0.24	102

Table 4: Classification report for the supervised model for inclusion and exclusion using a Logistic Regression classifier

Table 5 shows 18 hand-selected tweets as an example of how they are classified by the model and how in the original dataset. In tweets 1 and 3, it can be seen that despite the tweet containing the word ‘bevolkingsonderzoek’ (population study), the model correctly classified the tweets as ‘exclusion’. For tweet 2, the model also classified the tweet correctly, since this is about lung cancer and therefore not one of the cancer types being looked at for this study. In tweet 4, the tweet is also correctly classified as ‘exclusion’ by the model, as this is not a smear test for cancer. Tweets 5 to 8 are correctly classified as ‘inclusion’ by the model. In tweet 5 it is not literally named that it is about population screening for cancer but from the content it can be gathered that this is population screening for cervical cancer as it says ‘vijfjaarlijkse vrouwelijke checkup’ (five-yearly female check-up) and so this tweet indeed belongs to inclusion. In tweet 6, it is not named what the smear test is for, so this could be another doubtful case. But in the original dataset, this tweet is also classified as ‘inclusion’ so it can be assumed that the model got it right here too, based on the data the model was trained with. In tweet 7, it is again not written that it is population screening for cancer but again, the context classifies the tweet as ‘inclusion’. In tweet 8, it is named that it is population screening for colorectal cancer so again the model predicted it correctly. For tweets 9 to 13, the model did not perform well. Here, the model classified them as ‘inclusion’ while they were ‘exclusion’. Although it was said that when in doubt, tweets should be classified as ‘inclusion’ and are therefore not particularly severe errors, it is still interesting to see where it went wrong. Tweet 9 is about testing for an STI instead of cancer. Tweet 10 is about a different kind of population screening. Tweet 12 involves a Belgian tweet, of which it was indicated that it should be classified as ‘exclusion’. There will probably not have been many tweets in the original dataset where it was clear that it was a Belgian tweet, so it makes sense that the model finds it difficult to recognise them. Thereby, here it is indicated with an emoji that it is about Belgium. It is possible that that specific emoji was not part of one of the 1500 most important features of all the tweets and thus the model did not use it in its classification. Tweet 13 contains the word ‘smear’ but here it is not a smear for cancer screening but for pregnancy. For tweets 14 to 18, the model also did not perform well. Because of this error, tweets that are important for the study may be missed. Therefore, especially for this group, it is interesting to see where the model went wrong. Interestingly, tweets 14 and 15 have a hashtag (#) in front of all the words that could be key descriptors for the model. It could be that this caused the model not to recognise those tweets as inclusions. For tweets 16 to 17, it is more difficult to see why the model did not classify these tweets as inclusion.

Cancer Type In contrast to the inclusion model, the cancer type model demonstrates strong performance across all the different classes shown in Table 6. The ‘general or unclear’ class has the lowest score, which is expected as this category does not have distinct features that clearly define a specific cancer type. In comparison, the other classes have stronger key features, making them easier for the model to predict accurately, which will be discussed in Section 5.2. The precision and recall score of cervical and breast cancer shows that the model works well for identifying these classes. This high performance can be attributed to, not only the strong key features but also the dataset distribution, as these two classes were more present in the original dataset than the other classes. The ‘colorectal cancer’ class has the fewest cases in the original dataset. It still has a precision score of 1.00, indicating that the model labelled all the predicted

colorectal cancer cases correctly. However, with a recall score of 0.90, the model misses 10% of the actual colorectal cancer cases.

Cancer Type	Precision	Recall	Test instances
General or unclear	0.85	0.89	53
Cervical cancer	0.93	0.96	71
Breast cancer	0.96	0.95	78
Colorectal cancer	1.00	0.90	31

Table 6: Classification report for the supervised model for the category cancer type using a Random Forest classifier

Table 7 presents the performance of the Logistic Regression classifier for this category. While this classifier has lower overall precision and recall scores compared to the Random Forest classifier, the difference is not big enough to disregard it entirely. Therefore, it can still be used to gain more insight and be used for analysing the key features discussed in Section 5.2.

Cancer Type	Precision	Recall	Test instances
General or unclear	0.82	0.60	53
Cervical cancer	0.91	0.94	71
Breast cancer	0.80	0.95	78
Colorectal cancer	1.00	0.87	31

Table 7: Classification report for the supervised model for the category cancer type using a Logistic Regression classifier

Table 8 shows 12 tweets as examples that were classified the same in both the original dataset and the model. In tweets 1 and 2 the type of cancer is not explicitly mentioned and it cannot be identified from the context either, hence they were both categorized as ‘general or unclear’.

Tweets 3 to 5 are classified under ‘cervical cancer’. Tweet 3 contains the word ‘baarmoederhalskanker’ (cervical cancer) and thus belongs to the ‘cervical cancer’ class. For the other two tweets, it can be identified from the content. Tweet 4, mentions the word ‘uitstrijkje’ (smear), which indicates cervical cancer, while tweet 5 refers to the same term along with ‘30’, reinforcing the classification. Tweets 6 to 10 are classified as ‘breast cancer’. For tweets 6 to 8, this can again be identified from the context. Tweet 6 mentions ‘mammografie’ (mammography), tweet 7 ‘M’n tieten weer geplet’ (My boobs squashed again) and tweet 8 refers to being 50 and ‘memmenpletter’ (boobssquasher), all suggesting breast cancer. The classification of tweets 9 and 10 are notable. These were placed by both the original dataset and the model in the ‘breast cancer’ class, but if you look at the context, these tweets could have been placed under the ‘cervical cancer’ class. Breast cancer screening has a

two-yearly check-up, while this tweet uses the word ‘vijfjaarlijkse’ (five-yearly), which is how often cervical cancer screening occurs. Tweet 10 mentions almost turning 30, again this would rather point to cervical cancer as you get a call for cervical cancer screening at the age of 30, while breast cancer screenings typically start at 50. Neither tweets contain clear indicators of breast cancer, therefore it would be expected that these two tweets would be classified as cervical cancer rather than breast cancer in the original dataset. Since the model is trained with this dataset, this could potentially lead to misclassification in other tweets.

Tweets 11 and 12 are correctly classified as colorectal cancer, with both explicitly using the term ‘darmkanker’ (colorectal cancer). Tweet 11 only contains the word behind a hashtag #, which would mean that the model can also recognize some terms with a hashtag in front of it, although this does not always seem the case, as seen in Table 5.

Table 9 presents five tweets that the model misclassified. In tweet 1, the word ‘baarmoederhalskanker’ (cervical cancer) is explicitly mentioned, making it surprising that the model did not classify it correctly. In tweet 2, while the word ‘borstkanker’ (breast cancer) is not directly stated, the term ‘mammografie’ (mammography) is present, suggesting that it should have been classified under the breast cancer category. Tweet 3 contains the word ‘baarmoederhalskanker’ (cervical cancer), yet the model incorrectly classified it as breast cancer, which is again notable. In tweet 4, the model labelled the tweet as breast cancer, even though no specific type of cancer is mentioned, meaning it should have been classified as ‘general or unclear’. Conversely, tweet 5 should have been classified as ‘breast cancer’ but was instead labelled as ‘colorectal cancer’ by the model. Based on these examples alone, it is challenging to pinpoint why the model made these errors. In Section 5.2, the analysis of the most important features for each class will be conducted, which may provide insights into this misclassification.

Topic Overall, this model does not perform very well, which is shown in Table 10. Both the ‘legislation and regulations’ and ‘events’ classes have a precision score of 1.00, meaning that all the tweets labelled in this class by the model are correctly labelled, but the recall scores of 0.20 and 0.33, respectively, show that the model fails to correctly recognise a large proportion of tweets from these classes. This model performs especially poorly for the ‘science’ class, where both the precision and recall scores are 0.00, meaning that the model failed to correctly place any tweets in this class. The only class where the model performs reasonably is the ‘personal’ class, which is not surprising as this class has a distribution that is more than eight times larger than most other classes. Interestingly, the ‘other’ class relatively does not score very low, which might be unexpected as this category does not have such distinctive key features as the other classes, but is explainable since this class does have substantially more cases than the other categories, except for the ‘personal’ class.

The other model divided between the classes ‘personal’ and ‘not personal’ works substantially better which can be seen in Table 11. The difference in test instances is smaller which results in better recall and precision scores. With a recall score of 0.98 the model works well for identifying the personal cases.

Topic	Precision	Recall	Test instances
Other	0.53	0.25	32
Legislation and regulations	1.00	0.20	5
Events	1.00	0.33	6
Science	0.00	0.00	8
News	0.43	0.19	16
Personal	0.78	0.98	166

Table 10: Classification report for the supervised model for the category topic

Topic	Precision	Recall	Test instances
Not Personal	0.91	0.46	67
Personal	0.82	0.98	166

Table 11: Classification report for the supervised model for the category topic using a Random Forest classifier, simplified to a binary classification between the subcategories personal and not personal

Table 12 shows the performance of the Logistic Regression classifier for this category. Similar to the cancer type category, the Logistic Regression classifier shows lower overall precision and recall scores but remains sufficient for analysing the key features discussed in Section 5.2.

Topic	Precision	Recall	Test instances
Not Personal	0.81	0.43	67
Personal	0.81	0.96	166

Table 12: Classification report for the supervised model for the category topic using a Logistic Regression classifier, simplified to a binary classification between the subcategories personal and not personal

Table 13 shows 6 hand-selected tweets with their classification. In the original dataset, tweet 1 is classified as ‘news’ and tweet 2 as ‘other’. So, the model correctly classified both tweets as ‘non-personal’. Tweet 1 is non-personal as it contains news and no personal experiences. Tweet 2 I find harder to classify, which I have found with more tweets classified as ‘other’ in the original dataset. The tweet contains the word ‘ik’ (I) and is about receiving an invite for a smear test, which means it could also be seen as personal. However, the original dataset classified it as ‘other’ so the model correctly classified it.

Tweets 3 and 4 are about personally having participated in cancer screening and receiving the test

results. These tweets were correctly classified as personal. Tweets 5 and 6 were incorrectly classified as personal by the model. Tweet 5 could be seen as a personal opinion but it states in Section 3.2 that the contents of a personal opinion should be categorised. In the original dataset, this tweet was classified as ‘legislation and regulations’. Tweet 6 is classified as ‘other’ in the original dataset and I cannot identify a clear reason why the model assigned it to the personal category. It is possible that the word ‘ons’ (our) influenced this classification, or perhaps other words that serve as key features for the personal class, which I will discuss in Section 5.2.

Sentiment analysis Table 14 shows 7 hand-selected tweets with the sentiment assigned by the model, along with the certainty score of that sentiment. Tweets 1 to 3 have been assigned positive sentiment. I would personally also give tweet 1 a positive sentiment since it says the result has no abnormalities and because of the word “Top” (Great). The sentiment analysis for tweets 2 and 3 is a bit more complicated. Tweet 2 is probably a sarcastic comment but the model did not catch that and also gave it a fairly high certainty score of 0.86. The sentiment analysis of tweet 3 has been given a certainty score of 0.53, so the model did recognize that things are a bit more complicated with this tweet. The tweet is positive about the said article, but negative about population-based cancer screening.

Tweet 4 is about being up for population screening for breast cancer, there are no other opinions attached to this tweet and so the model correctly classified it as neutral.

Tweets 5 to 7 have been assigned a negative sentiment. Tweet 5 is about how the population screening hurt and it has been assigned a certainty score of 0.88. Tweet 6 is about getting the results of the screening, in which fortunately nothing was found. By the word ‘gelukkig’ (fortunately), I would think that this one should have been assigned a positive sentiment. This one has also been given a slightly lower certainty score of 0.53. Tweet 7 is also interesting, with a certainty score of 0.39. This tweet again seems to involve a sarcastic remark. Unlike tweet 2, the model seems to have recognised it here, but the low certainty score shows that the model still finds it tricky.

As mentioned in Section 4.1, I also tested a Dutch pre-trained model from Hugging Face. The results of this model are shown in Table 15. One limitation of this model which I already mentioned is its inability to classify tweets with a neutral sentiment, like the chosen model did for tweet 5. Additionally, I found its overall performance not as good as the performance shown in Table 14. Tweets 5 and 8 were classified as positive by the Dutch model, while I believe the other model classified it more accurate in these cases. Notably, both models classified tweet 6 as negative, while I would have classified it as positive. However, the chosen model did so with a certainty score of 0.53, while the Dutch model classified it with a score of 1.00. Based on these observations, I found the other model to perform better overall than this Dutch model.

5.2 Key features

Cancer type As discussed in Section 4.2 the Logistic Regression classifier can provide more insight into the key features of each cancer type than the Random Forest classifier. The Tables 16, 17, 18 and 19 show the key features of each cancer type where the logistic regression model was used described in Section 4.2.

The ‘general or unclear’ class has on average the lowest importance scores. The words ‘bevolkingsonderzoek’ (population screening) and ‘kankerscreening’ (cancer screening) are the most important features with still relatively high feature importance scores. However, compared to the top 10 in the other classes, the importance scores for this class tend to be lower. This shows that the most relevant features for this class are less distinctive, making it harder for the model to classify tweets into this category confidently.

For the other three classes, the name of the type of cancer and the test to detect it are the two most important key features. Interestingly, in the cervical cancer class, the key feature ‘uitstrijkje’ (smear) has a higher importance score than the word ‘baarmoederhalskanker’ (cervical cancer) itself. Similarly, for the breast cancer class, the word ‘mammografie’ (mammography) has a higher importance score than ‘borstkanker’ (breast cancer). For the colorectal cancer class, the word ‘darmkanker’ (colorectal cancer) does have a higher importance score than ‘ontlastingstest’ (stool test).

It is also notable that the number 30 has a high importance score in cervical cancer, while the ages of 50 in breast cancer and 55 in colorectal cancer do not. In breast cancer, the number 50 has an importance score of 0.57 and ranks 18 among the key features. The number 55 in colorectal cancer has an importance score of 0.39 and ranks at spot 40. A potential reason for this is that Twitter users might tweet that they turned 30 and received their first invite for a population-based cancer screening. There are also more Dutch Twitter users around the age of 30 than the age of 50 or 55, according to a study conducted in 2013 [NGTM13].

Another interesting point is that in the cervical cancer class, the words ‘een’ (a(n)), ‘geen’ (none), ‘als’ (as/if) and ‘je’ (you) score high as key features with relatively high importance scores compared to the scores of the key features in the other classes and compared to words that would seem more meaningful as a key feature.

Another word with a high importance score for the cervical cancer class is ‘ik’ (I). This suggests that people may share more personal experiences when discussing cervical cancer compared to the other types of cancer. In contrast, the word ‘mensen’ (people) appears in the top 10 for the colorectal cancer class, which could indicate that conversations about this type of cancer on social media are generally more broad and impersonal.

The 10 key features of the breast cancer class mainly consist of various terms related to the screening test and different words for breasts. A notable key feature is the word ‘ziekenhuis’ (hospital). This is interesting compared to the key feature ‘huisarts’ (GP) for the cervical cancer class. This comparison highlights that for both classes, the location where the test is conducted is also relevant.

For colorectal cancer, the words ‘bevolkingsonderzoek’ (population screening) and ‘screening’ (screening) rank relatively high in the feature importance list, even though these words do not point to colorectal cancer specifically. The year 2013 is also high on the list here. It was planned to begin the population screening for colorectal cancer in 2013 [RIV12].

For breast cancer and cervical cancer, it was as early as 1990 and 1996, respectively.¹² ¹³ Since the data used for this study contains tweets from 2010 onwards, it makes sense that 2013 is a key feature here and the other year numbers are not. The word ‘start’ (start) is also high on the list, which could be referring to the start of the population-based colorectal cancer screening in 2013 which also could give a reason for the high scores of the words ‘bevolkingsonderzoek’ (population screening) and ‘screening’ (screening).

Topic Table 20 shows that ‘darmkanker’ (colorectal cancer) is the top key feature for the non-personal class. This indicates that conversations on social media regarding colorectal cancer are indeed generally more broad and impersonal. Table 21 highlights that, as expected, ‘ik’ (I) is the top key feature for the personal class. Additionally, ‘uitslag’ (result) ranks as the second most important key feature, suggesting that when people share personal experiences, they often mention receiving or awaiting their results. Combined with the key feature ‘goed’ (good) this could imply that individuals frequently share that they have received good test results from cancer screening. In Section 5.1 I discussed Table 13 which shows that tweets 5 and 6 from that table were incorrectly classified as ‘personal’. Initially, I assumed this could have been due to the tweets containing key features for this class. However Table 20 shows that tweet 5 includes two words – ‘laat’ (let) and ‘vrouwen’ (women) – that are in the top 5 key features for the non-personal class. Similarly, tweet 6 contains the word ‘darmkanker’ (colorectal cancer) which is the top key feature for the non-personal class. Neither tweets contain any words from the top 5 key features for the personal class. Therefore, the top 5 key features from both the personal and non-personal classes cannot explain why the model misclassified these tweets as ‘personal’.

5.3 Analysis of the large dataset

Inclusion Figure 5 shows the distribution of the exclusion and inclusion classes for both the original dataset and the model’s classification of all tweets. It’s important to keep in mind that the dataset the model classified is three times larger than the original dataset. Despite this difference in size, the original dataset has more tweets classified as exclusion compared to the larger dataset classified by the model. This matches with what is described in Section 5.2, where it is explained that the model misses many true exclusion cases, which is reflected in this Figure.

¹²<https://www.rivm.nl/media/cvb/25jaarbvbk/index.html>

¹³<https://www.rivm.nl/bevolkingsonderzoek-baarmoederhalskanker/professionals/achtergrond>

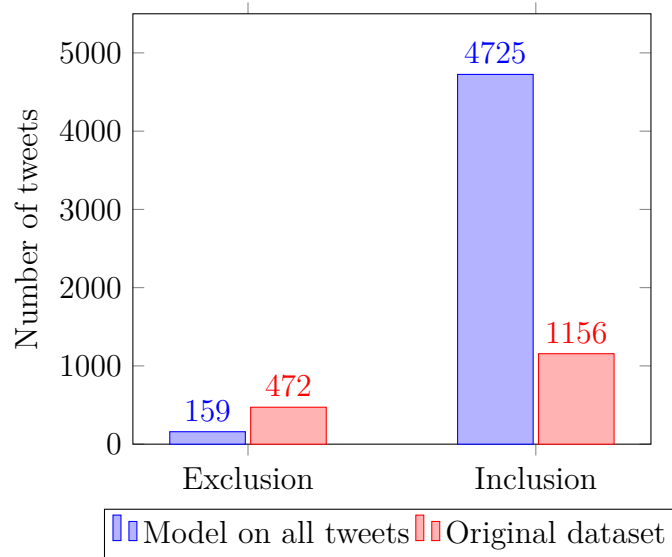


Figure 5: Distribution exclusion and inclusion for the results of the supervised model and the original labelled dataset

Cancer type Figure 6 shows the distribution of classifications for the different types of cancer. Here, the dataset of all tweets is more than four times larger than the original dataset, since the exclusion cases are not part of the cancer type classifications in the original dataset. It’s interesting that in the original dataset, tweets are classified as ‘general or unclear’ much more often than by the model when applied to all tweets. As shown in Table 6, the ‘general or unclear’ class has a recall of 0.89. While this is lower than the other classes, it is not so low that it would fully explain the smaller number of tweets classified as ‘general or unclear’ by the model. This may be due to the ‘general or unclear’ class lacking strong, distinctive key features, as discussed in Section 5.2. Another noticeable point is that the model classifies tweets as ‘breast cancer’ much more often than in the original dataset. This difference could be because of possible mistakes in the original dataset, as discussed in Section 5.1 with Table 8, where tweets 9 and 10 might have been incorrectly labelled as ‘breast cancer’. Such mistakes in the training data could influence how the model learns, making it classify more tweets as ‘breast cancer’ which might explain the higher number seen in this category.

Topic Figure 7 presents the distribution of ‘personal’ versus ‘not personal’ classifications. This distribution aligns most closely with the original dataset compared to the other classifications. However, the model has labelled a relatively higher number of tweets as ‘personal’ than the original dataset. This outcome is consistent with the findings in Section 5.1, where the recall for the ‘not personal’ class is 0.46, indicating that many actual ‘not personal’ cases were missed. As a result, the higher proportion of ‘personal’ classifications in this distribution is not unexpected.

Sentiment analysis Figure 8 shows the distribution of the sentiment analysis results. The majority of tweets are classified as neutral, with 1912 tweets, followed by 1673 negative tweets, and 1299 positive tweets. Given that the focus of this study is to understand why some individuals do not respond to the invitation for population screening programmes for cancer and how social media

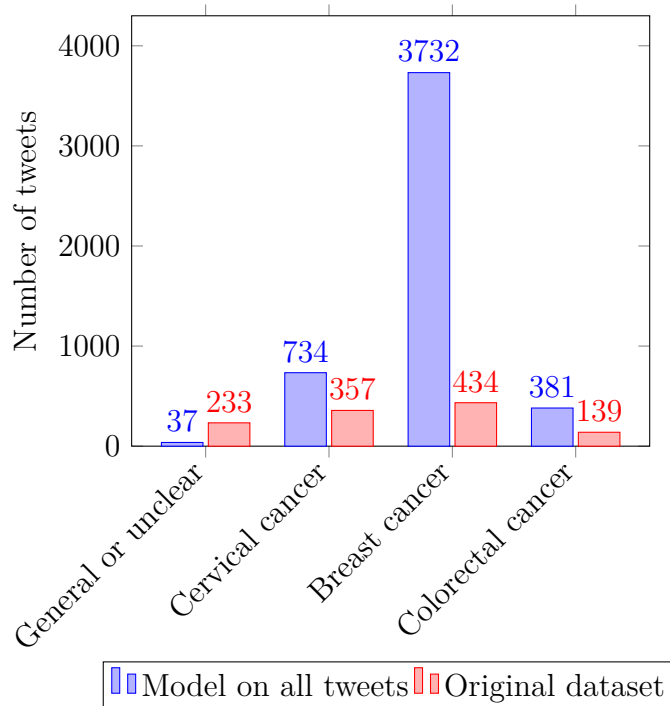


Figure 6: Distribution for the results of the supervised model and the original labelled dataset for the category cancer type

might influence this decision, it is particularly noteworthy that such a large number of tweets are classified as negative.

Cross tables Table 22 shows a cross table comparing cancer type and sentiment, while Table 23 compares topic and sentiment. Both tables do not show any notable findings. It is expected that neutral scores the highest in each subcategory as the overall sentiment distribution is also highest for neutral. For the colorectal cancer class and the non-personal class, the proportion of negative sentiment is higher than neutral, but this is not a substantial difference. I expected the non-personal class to have a larger proportion of neutral sentiment as this subcategory does not include positive or negative experiences, which are likely to be connected to the given sentiment. Another interesting observation is that the neutral sentiment has the highest distribution in the general or unclear class for cancer type. This may indicate that tweets in this subcategory, being less specific and harder to classify, also have a vaguer sentiment, which could explain the higher distribution of neutral sentiment in this class. Another potential explanation could be that news reports, which are more likely to have a neutral sentiment, might be more often about cancer screening in general than about a specific cancer type.

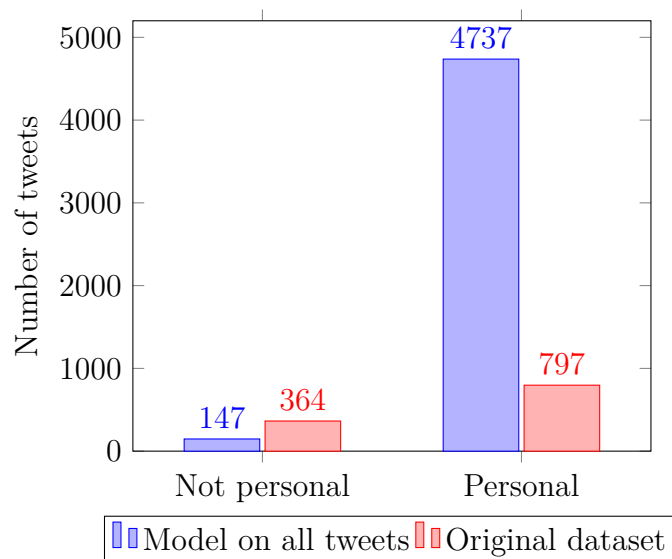


Figure 7: Distribution for the results of the supervised model and the original labelled dataset for the category topic

6 Discussion

Based on the results in Section 5.1, I can state that the model can be used for pre-filtering tweets to determine which ones should be included because the model can successfully find almost all the inclusion cases. It does classify a substantial exclusion cases as inclusion, but for pre-filtering that is okay and as it was stated by the SENTENCES project that, when in doubt, the tweets should be categorised as inclusion. It is important to note that while the model works sufficiently on this dataset, its performance may not be the same on other datasets. This is due to the selection bias of the data, as it was filtered on the keywords described in Section 3.1.

The results also show that the model can be used for categorising the tweets for the different cancer types. This model works well as each category has some distinct key features, except for the ‘General or unclear’ class but even there the model shows sufficient recall and precision scores. What is interesting is that the model classifies tweets as ‘breast cancer’ much more often than in the original dataset. As discussed in Section 5.3 this could have been caused by mistakes in the training data. The model for topic is less useful than the other two supervised models, because of the big imbalance.

For all three categories there are tweets misclassified by the model, for which I could not find an explanation. The key features account for some (mis)classifications. Additionally, it is possible that hashtags and emojis influenced the results, as they occur less frequently than other important words and therefore may have given a much lower importance score during the vectorizing process or even not been included in the top 1000 or 1500 features. As a result, they could have not been considered by the model, even though hashtags and emojis often convey the essence of what the writer intends to express in a tweet, which is why hashtag analysis could be interesting.

It is also worth noting that the models were trained on approximately 1303 tweets (80% of 1629), which is relatively limited compared to the datasets typically used to train models, such as the pre-trained model I used for sentiment analysis which was trained on around 198 million tweets.

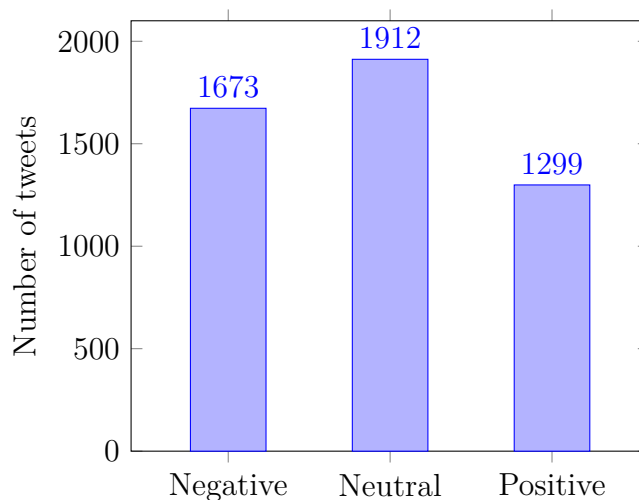


Figure 8: Distribution for the results of the unsupervised model for sentiment

Something notable about the analysis of the key features is that the key features for both the cancer type category and the topic category indicate that discussions on social media about colorectal cancer are more general rather than personal. Sentiment analysis can be useful as it can give a good overview of the distribution of sentiment. Something to consider is that the model judges the sentiment of the whole tweet, the tweet may be positive but talk negatively about cancer screening, or vice versa. In some cases, I do not agree with the sentiment assigned by the model, but often, I am also unsure about the sentiment I would assign to the tweet myself. I would recommend using the sentiment classification provided by the model when the sentiment score is higher than 0.70. Otherwise, I would advise to manually review the sentiment.

7 Conclusions and further research

The aim of this research was to address the question ‘How can we use automatic text analytics methods to identify and categorise narratives about cancer screening?’. Through this study, I have found that the use of automatic text analytics methods can assist in identifying and categorising narratives about cancer screening. To answer the question ‘How well can automatic text analytics methods classify social media posts about cancer screening?’ I have found that the overall high recall scores of the models indicate that they successfully classify many tweets correctly, though some instances are still missed. While these models can serve as a helpful tool for filtering content, it is necessary for humans to manually review the results to ensure accuracy. For large datasets, these models can provide a reasonably accurate overview of the data. Sentiment analysis can offer further insights.

The analysis of key features for the different categories reveal the answer to the question ‘Which words serve as key descriptors for categorising these posts?’. The ‘general or unclear’ category displayed lower average importance scores, with the words ‘bevolkingsonderzoek’ (population screening) and ‘kankerscreening’ (cancer screening) leading the list. For each cancer type, the most important descriptors included the name of the cancer type and the associated screening test. For the classification on the topic category, the words ‘ik’ (I) and ‘uitslag’ (result) were the key

descriptors for the personal class and the words ‘darmkanker’ (colorectal cancer) and ‘screening’ (screening) for the non-personal class.

To answer the last sub-question ‘What is the distribution of these categories within a large dataset of 4885 social media posts?’, I looked at the distribution within the complete dataset which shows that, among the discussed types of cancer, cancer screening for breast cancer is most mentioned on social media. The distribution further indicates that, in general, cancer screening is discussed more negatively than positively on social media. There is also no substantial difference found in how positive or negative a certain subcategory is discussed compared to others.

For further research, I would recommend looking at a new sample of the total 4885 tweets and see if the distribution of the breast cancer class is indeed relatively higher than what it seems in the subset of 1629 tweets. Otherwise, it could mean that there were indeed mistakes in the training data that led to the model incorrectly classifying tweets as ‘breast cancer’. And then the model should be re-trained on more accurate data. Furthermore, I would use the same steps described in Section 4 to make models for the other categories mentioned in Section 3.2. I would also look at hashtag analysis, for further analysis of narratives about cancer screening on social media.

References

- [BBC⁺22] Thomas HG Bongaerts, Frederike L Büchner, Matty R Crone, Job van Exel, Onno R Guicherit, Mattijs E Numans, and Vera Nierkens. Perspectives on cancer screening participation in a highly urbanized region: a q-methodology study in the hague, the netherlands. *BMC public health*, 22(1):1925, 2022.
- [BMC21] Priscila Biancovilli, Lilla Makszin, and Alexandra Csongor. Breast cancer on social media: a quali-quantitative study on the credibility and content type of the most shared news stories. *BMC Women’s Health*, 21(1):202, 2021.
- [CCJ⁺16] W Christian Crannell, Eric Clark, Chris Jones, Ted A James, and Jesse Moore. A pattern-matched twitter analysis of us cancer-patient sentiments. *Journal of Surgical Research*, 206(2):536–542, 2016.
- [CJJ⁺18] Eric M Clark, Ted James, Chris A Jones, Amulya Alapati, Promise Ukandu, Christopher M Danforth, and Peter Sheridan Dodds. A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. *arXiv preprint arXiv:1805.09959*, 2018.
- [CLC23] Jefferson C Chen, Christina A LeBedis, and Kevin J Chang. The public perception of ct colonography versus colonoscopy via sentiment analysis of social media. *Journal of the American College of Radiology*, 20(6):531–536, 2023.
- [CWP18] Liang Chen, Xiaohui Wang, and Tai-Quan Peng. Nature and diffusion of gynecologic cancer–related misinformation on social media: analysis of tweets. *Journal of Medical Internet Research*, 20(10):e11515, 2018.
- [EODLILE20] Oduwa Edo-Osagie, Beatriz De La Iglesia, Iain Lake, and Obaghe Edeghere. A scoping review of the use of twitter for public health research. *Computers in biology and medicine*, 122:103770, 2020.

- [HVdBD23] Hanneke Hendriks, Suzan Verberne, Gert-Jan de Bruijn, and Enny Das. (wip) news coverage about cancer screening in the netherlands: A content analysis: Presenter (s): Inge stortenbeker, radboud university, netherlands. *Patient Education and Counseling*, 109:62–63, 2023.
- [JCWG15] Xiang Ji, Soon Ae Chun, Zhi Wei, and James Geller. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5:1–25, 2015.
- [KLL17] S. Koval, Y. Li, and M. Lyst. A big data challenge: Visualizing social media trends about cancer using sas® text miner. *Pinnacle Solutions, Inc.*, 2017.
- [KTK+23] Nari Kureyama, Mitsuo Terada, Maho Kusudo, Kazuki Nozawa, Yumi Wanifuchi-Endo, Takashi Fujita, Tomoko Asano, Akiko Kato, Makiko Mori, Nanae Horisawa, et al. Fact-checking cancer information on social media in japan: Retrospective study using twitter. *JMIR Formative Research*, 7(1):e49452, 2023.
- [LRL+19] Gem M Le, Kate Radcliffe, Courtney Lyles, Helena C Lyson, Byron Wallace, George Sawaya, Rena Pasick, Damon Centola, and Urmimala Sarkar. Perceptions of cervical cancer prevention on twitter uncovered by different sampling strategies. *PloS one*, 14(2):e0211931, 2019.
- [MBLS17] Omar Metwally, Seth Blumberg, Uri Ladabaum, and Sidhartha R Sinha. Using social media to characterize public sentiment toward medical interventions commonly used for cancer screening: an observational study. *Journal of medical Internet research*, 19(6):e200, 2017.
- [Ned22] Integraal Kankercentrum Nederland. National monitoring of the breast cancer screening programme in the netherlands 2020/2021, 2022.
- [Ned24] Integraal Kankercentrum Nederland. Cancer screening, 2024.
- [NGTM13] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. ” how old do you think i am?” a study of language and age in twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 439–448, 2013.
- [POP+16] SoHyun Park, Heung-Kwon Oh, Gibeom Park, Bongwon Suh, Woo Kyung Bae, Jin Won Kim, Hyuk Yoon, Duck-Woo Kim, and Sung-Bum Kang. The source and credibility of colorectal cancer information on twitter. *Medicine*, 95(7):e2775, 2016.
- [RIV12] RIVM. Bevolkingsonderzoek darmkanker vanaf 2013, 2012.
- [SGP+18] Rosa Sicilia, Stella Lo Giudice, Yulong Pei, Mykola Pechenizkiy, and Paolo Soda. Twitter rumour detection in the health domain. *Expert Systems with Applications*, 110:33–40, 2018.

- [TSV⁺18] Deanna Teoh, Rida Shaikh, Rachel Isaksson Vogel, Taylor Zoellner, Linda Carson, Shalini Kulasingam, and Emil Lou. A cross-sectional review of cervical cancer messages on twitter during cervical cancer awareness month. *Journal of lower genital tract disease*, 22(1):8–12, 2018.








	Tweet	model	original dataset
1	Bevolkingsonderzoek Q-koorts valt in de categorie domme plannen van mensen die er niets van begrepen hebben	exclusion	exclusion
2	Met 20.000 tabaksdoden elk jaar in Nederland en oorzaak 30% alle kanker sterfte hoort bij elke screening een rookstopadvies. Meisjes van 13	exclusion	exclusion
3	komt er nog een groot bevolkingsonderzoek rond harlingen en het waddengebied daar? omrin: hoeft niet, politiek: is het wel nodig? zo duur.	exclusion	exclusion
4	Jammer, de echo levert te weinig op. Gelijk uitstrijkje gemaakt. Over 2 weken naar ziekenhuis voor watercontrastecho om te zien hoe of wat	exclusion	exclusion
5	Bij de huisarts, de vijfjaarlijkse vrouwelijke checkup #blech . Straks is het weer achter de rug #bevolkingsonderzoek	inclusion	inclusion
6	In de wachtkamer voor een uitstrijkje... Gezellig!	inclusion	inclusion
7	Jee, nou ben ik ook al oud genoeg voor het bevolkingsonderzoek.....dan ben je dus echt oud geworden.....	inclusion	inclusion
8	...'screening werkt' staat er in de krant. ' ...minder darmkanker door groot bevolkingsonderzoek' ...euh, wat mis ik hier ?	inclusion	inclusion
9	Met 'n #zelfafnametest vr #Chlamydia en #Gonorroe v SoaCare ben je verzekerd v dezelfde #betrouwbaarheid als v de huisarts http://soacare.nl	inclusion	exclusion
10	Er stond een damesfiets bij de bus van het bevolkingsonderzoek.	inclusion	exclusion
11	Vanmorgen weer mammogram laten doen. Pfff altijd spannend. 2x overnieuw. Extra spanning. Maar gelukkig weer GOED. Over een jaar maar were.	inclusion	exclusion
12	 (1/7) In  krijgen jaarlijks 2000 mensen een "vroegtijdig" stadium II darmkanker. Dit aantal is in stijgende lijn dankzij het bevolkingsonderzoek  . Ongeveer 1/3 van hen zal na de operatie worden behandeld met een preventieve chemotherapie   om de genezingskans te verhogen	inclusion	exclusion
13	#moe in december blijkbaar een bacterie opgelopen tijdens de zwangerschap, erachter gekomen met een uitstrijkje, wat kan je van een zware antibiotica beroerd worden zeg	inclusion	exclusion
14	Gaat het #bevolkingsonderzoek naar #darmkanker nog door @rivm @MLDS @MinVWS ? #codezwart	exclusion	inclusion
15	Testuitslag: negatief   #bevolkingsonderzoek #darmkanker	exclusion	inclusion
16	Operatiekamers ADRZ in Goes moeten state of the art vanwege onder meer bevolkingsonderzoek darmkanker. Zal meer patiënten opleveren #adrzoz	exclusion	inclusion
17	net #overmijlijk gekeken en zometeen naar de huisarts voor een uitstrijkje.... besef me na deze aflevering maar des te meer weer hoe belangrijk dat is.	exclusion	inclusion

Table 5: Tweets with the classification of the supervised model and the original labelled dataset for the category inclusion

	Tweet	Cancer type
1	#LETBSymposium luisteren naar spreekster Nynke de Jong @Rivm. Over bewezen effect van bevolkingsonderzoek	general or unclear
2	Maak morgen kennis met het team van @vigezine achter #gezondegemeente op #gezondheidsconferentie #bevolkingsonderzoek Tot morgen.	general or unclear
3	"Een oproep voor bevolkingsonderzoek baarmoederhalskanker en je belt voor een afspraak. Kan over 3 maanden terugbellen want wordt nu niet gedaan vanwege drukte in de praktijk. Ik vind er iets van..."	cervical cancer
4	RIVM: jaarlijks tientallen doden door negeren oproep uitstrijkje ... Om half 12 op @NPORadio1 met afdelingshoofd Nynke van der Veen @rivm en @SPnl @HenkvGerven	cervical cancer
5	"Ben over 1 week pas 30 maar het de papieren voor het bevolkingsonderzoek al binnen. Dinsdag een uitstrijkje brrrrrrrrrrrr."	cervical cancer
6	BlvdD 12 mrt:voor het eerst een mammografie, het viel me alles mee. De vrouw die het doet -ik zeg, wat een rare baan is dit- is blij met hoeveel vrouwen ze redt door een vroege diagnose. #bevolkingsonderzoek	breast cancer
7	Oproep bevolkingsonderzoek. M'n tieten weer geplet! :/	breast cancer
8	Jaaaaa hoor! Ben je koud 50 of daar ligt een uitnodiging voor de menmenpletter op de mat! #bevolkingsonderzoek	breast cancer
9	Bij de huisarts, de vijfjaarlijkse vrouwelijke checkup #blech . Straks is het weer achter de rug #bevolkingsonderzoek	breast cancer
10	Zojuist geconfronteerd met dat ik bijna 30 ben. Tijdschrift met uitleg over #bevolkingsonderzoek op de mat.	breast cancer
11	Als we sterfte aan darmkanker willen terugdringen, en wie wil dat niet, dan is screening/ #darmkankeronderzoek het enige realistische middel	colorectal cancer
12	Bevolkingsonderzoek naar #darmkanker gaat geleidelijk van start. Maandag in #Ommelander Courant.	colorectal cancer

Table 8: Tweets with the correct classification of the supervised model for the category cancer type

	Tweet	model	original dataset
1	Wat verandert er vanaf 2017 in bevolkingsonderzoek baarmoederhalskanker? Kom naar onze stand D073 #hab2016 #huisartsenbeurs	general or unclear	cervical cancer
2	"Vervelend wel dat ik gisteren te horen kreeg via mijn huisarts dat er iets was met een mammografie van anderhalf jaar geleden en dat ik een jaar geleden een nieuwe oproep had moet krijgen. #coronafoutjebedankt Nu woensdag AvL."	cervical cancer	breast cancer
3	Doe je mee aan het bevolkingsonderzoek baarmoederhalskanker. Krijg je een slechte uitslag. Kun je pas 6(!) weken later terecht voor verder onderzoek. Wat is er toch veel mis met onze zorg.	breast cancer	cervical cancer
4	O ja, bevolkingsonderzoek op de mat.	breast cancer	general or unclear
5	Net discussie met balletlesmoeders over mammografie (wat is erger:kleine borsten en dan gezeur dat t niet haalt, of grote :dat t niet past	colorectal cancer	breast cancer

Table 9: Tweets with the incorrect classification of the supervised model for the category cancer type

	Tweet	model	original dataset
1	Kwalijk gevolg van landelijke darmkanker screening: wachttijden bij MDL artsen van 45+ dagen voor mensen zonder screening maar met klachten!	Non-personal	Non-personal
2	Ik ken mijn plek weer. 1 brief voor een uitstrijkje, 1 rekening en 1 brief van de belastingdienst voor de burens. Lekker is dat #valentijn	Non-personal	Non-personal
3	goeiemorgen. Wat is t glad. Heb huisarts gebeld, pap1, dus goed #uitstrijkje. Er waren dus cervixcellen gevonden #blij	Personal	Personal
4	Net bij de dokter geweest voor het bevolkingsonderzoek, zomaar klaar. Hope dat de uitslag via de post komt. Nu eerst koffie #lekkerlekker	Personal	Personal
5	Laat vrouwen vanaf 30 jaar deelnemen aan een bevolkingsonderzoek naar borstkanker.Zou een hoop verschil maken,het zou eerder ontdekt worden.	Personal	Non-personal
6	in ons programma @isala Astrid Nauta over de publieksacademie 14 Maart 2016 landelijk bevolkingsonderzoek darmkanker en nieuwe technieken.	Personal	Non-personal

Table 13: Tweets with the classification of the supervised model and the original labelled dataset for the category topic

	Tweet	Sentiment	Score
1	Het resultaat van mijn #bevolkingsonderzoek #dikkedarmjanker is niet afwijkend. Top! #laatjetesten #stopdarmkanker	Positive	0.83
2	Leuk hoor, 30 worden. Of ik even een uitstrijkje wil laten maken.	Positive	0.86
3	Sterk artikel in #Trouw ‘Screening darmkanker biedt valse beloften’ van @caseofdees en @aliettejonkers #preventieparadox	Positive	0.53
4	Op 25 maart ben ik weer aan de beurt voor het @BVOZuidWest. #bevolkingsonderzoek #borstkanker	Neutral	0.87
5	GVD dat deed pijn #bevolkingsonderzoek	Negative	0.88
6	Vandaag uitslag gekregen van het bevolkingsonderzoek borstkanker..... gelukkig niets gevonden.	Negative	0.53
7	leuk’ verjaardagscadeau! uitnodiging (!) voor bevolkingsonderzoek baarmoederhalskanker #noodzakelijk #tandenopelkaar	Negative	0.39

Table 14: Tweets with the sentiment analysis with certainty scores of the unsupervised model

	Tweet	Sentiment	Score
1	Het resultaat van mijn #bevolkingsonderzoek #dikkedarmjanker is niet afwijkend. Top! #laatjetesten #stopdarmkanker	Positive	1.00
2	Leuk hoor, 30 worden. Of ik even een uitstrijkje wil laten maken.	Positive	0.98
3	Sterk artikel in #Trouw ‘Screening darmkanker biedt valse beloften’ van @caseofdees en @aliettejonkers #preventieparadox	Positive	1.00
4	Op 25 maart ben ik weer aan de beurt voor het @BVOZuidWest. #bevolkingsonderzoek #borstkanker	Positive	1.00
5	GVD dat deed pijn #bevolkingsonderzoek	Positive	0.51
6	Vandaag uitslag gekregen van het bevolkingsonderzoek borstkanker..... gelukkig niets gevonden.	Negative	1.00
7	leuk’ verjaardagscadeau! uitnodiging (!) voor bevolkingsonderzoek baarmoederhalskanker #noodzakelijk #tandenopelkaar	Positive	0.99

Table 15: Tweets with the sentiment analysis with certainty scores of the unsupervised model using the Dutch pre-trained model

Feature	Feature translated	Importance score
bevolkingsonderzoek	population screening	3.08
kankerscreening	cancer screening	1.23
dat	that	1.01
van	from	0.77
kanker	cancer	0.76
laat	let/late	0.73
huisartsen	GPs	0.71
afspraak	appointment	0.63
weken	weeks	0.63
leuk	nice	0.63

Table 16: 10 key features for the general or unclear class of the category cancer type

Feature	Feature translated	Importance score
uitstrijkje	smear	6.58
baarmoederhalskanker	cervical cancer	4.37
hpv	hpv	1.63
30	30	1.48
huisarts	GP	1.33
ik	I	1.20
een	a(n)	1.09
geen	no/none	0.90
als	as/if	0.83
je	you	0.79

Table 17: 10 key features for the cervical cancer class of the category cancer type

Feature	Feature translated	Importance score
mammografie	mammography	5.78
borstkanker	breast cancer	5.04
mammogram	mammogram	1.69
borsten	breasts	1.03
echo	echo	0.83
borstonderzoek	breast screening	0.81
tieten	boobs	0.77
is	is	0.74
ziekenhuis	hospital	0.66
borst	breast	0.64

Table 18: 10 key features for the breast cancer class of the category cancer type

Feature	Feature translated	Importance score
darmkanker	colorectal cancer	6.48
ontlastingstest	stool test	1.00
start	start	0.99
2013	2013	0.96
bevolkingsonderzoek	population screening	0.86
darm	intestine	0.83
screening	screening	0.76
mdl	stomach, intestine, liver	0.76
mensen	people	0.68
luchtje	smell	0.68

Table 19: 10 key features for the colorectal cancer class of the category cancer type

Feature	Feature translated	Importance score
darmkanker	colorectal cancer	2.73
screening	screening	2.24
vrouwen	women	1.22
start	start	1.19
laat	let/late	1.16

Table 20: 5 key features for the non-personal class of the category topic

Feature	Feature translated	Importance score
ik	I	2.59
uitslag	result	1.85
weer	again	1.76
goed	good	1.46
mammografie	mammography	1.42

Table 21: 5 key features for the personal class of the category topic

	Negative	Neutral	Positive
General or unclear	32%	54%	14%
Cervical cancer	33%	41%	26%
Breast cancer	34%	39%	27%
Colorectal cancer	40%	37%	23%

Table 22: Cross table for the results of the models for cancer type with sentiment

	Negative	Neutral	Positive
Non-personal	38%	33%	29%
Personal	34%	39%	27%

Table 23: Cross table for the results of the models for topic with sentiment

	Text	Inclusion	Cancer type	Topic
1	So painful #mammography not just my breast but also my neck had to be included. Was told not to exaggerate! Had large bruises for 3 weeks and the test failed. It has to be done, but I don't know how. @bvo_nederland	1	2	5
4	"After the dentist (Monday) and the booster shot (Tuesday), I received an invitation for the breast cancer screening programme on Wednesday 😞 Will you cross your fingers for me on 11 November in the afternoon? 😞	1	2	5
7	"The misery that is #Tinder: When you call your GP for an STD test... And they need to check if you also need a smear test. #awkward but so sensible."	0		
10	"Soon women aged 30+ will receive a kit to make their own smear test (cervical cancer). I'm trying to imagine that, but honestly, it doesn't seem that simple."	1	1	1

Table 24: Sample of translated tweets collection with labels

	Text
1	So painful #mammography not just my breast but also my neck had to be included. Was told not to exaggerate! Had large bruises for 3 weeks, and the test failed. It has to be done, but I don't know how. @bvo_nederland
2	Next week, my man (to call him that for once) has to go to the hospital. Photos of his "chest" – can men also have something wrong there? And just now, mail about the population screening, an appointment for a colonoscopy. Gross.
3	I have received two invitations for the breast cancer screening programme. Do I go with one first and then the other..?
4	"After the dentist (Monday) and the booster shot (Tuesday), on Wednesday came an invitation for the breast cancer screening programme 😞 Will you cross your fingers for me on 11 November in the afternoon? 😞 "

Table 25: Sample of translated large tweet collection without labels








	Tweet	model	original dataset
1	Population screening for Q fever falls into the category of dumb plans made by people who don't understand a thing about it	exclusion	exclusion
2	With 20,000 tobacco deaths every year in the Netherlands and 30% of all cancer deaths caused by smoking, every screening should include smoking cessation advice. Girls aged 13	exclusion	exclusion
3	is there going to be a large population screening around Harlingen and the Wadden area? omrin: no need, politics: is it necessary? So expensive.	exclusion	exclusion
4	What a shame, the ultrasound yields too little. Immediately made a smear test. In 2 weeks to the hospital for a water contrast ultrasound to see what's going on	exclusion	exclusion
5	At the GP's, the five-yearly female checkup #blech. Soon it will be over #population-screening	inclusion	inclusion
6	In the waiting room for a smear test... So much fun!	inclusion	inclusion
7	Jeez, now I'm old enough for the population screening.....so I'm really getting old.....	inclusion	inclusion
8	...'screening works' is written in the newspaper. '...less colorectal cancer due to large population screening' ...uh, what am I missing here ?	inclusion	inclusion
9	With a #self-sampling test for #Chlamydia and #Gonorrhoea from SoaCare, you're guaranteed the same #reliability as from the GP HTTP://soacare.nl	inclusion	exclusion
10	There was a women's bike parked by the population screening van.	inclusion	exclusion
11	This morning had another mammogram done. Phew, always nerve-racking. Twice over. Extra stress. But luckily, all GOOD again. Back in a year.	inclusion	exclusion
12	 (1/7) In  , every year 2,000 people are diagnosed with "early-stage" stage II colorectal cancer. This number is increasing thanks to the population screening  . About 1/3 of them will be treated with preventive chemotherapy   after surgery to increase their chances of recovery.	inclusion	exclusion
13	#tired in December, apparently caught a bacterial infection during pregnancy, found out with a smear test, wow antibiotics can make you feel awful.	inclusion	exclusion
14	Will the #population screening for #colorectal cancer continue @rivm @MLDS @MinVWS? #codeblack	exclusion	inclusion
15	Test result: negative   #population-screening #colorectal cancer	exclusion	inclusion
16	Operating rooms at ADZ in Goes need to be state of the art due to, among other things, population screening for colorectal cancer. It will bring in more patients #adrzoz	exclusion	inclusion
17	Just watched #overmijnlijk and now heading to the GP for a smear test.... After this episode, I realise once again how important it is.	exclusion	inclusion

Table 26: Translated tweets with the classification of the supervised model and the original labelled dataset for the category inclusion

	Tweet	Cancer Type
1	#LETBSymposium listening to speaker Nynke de Jong @Rivm. About the proven effect of population screening	general or unclear
2	Meet the team behind @vigezine tomorrow at #healthymunicipality during #healthconference #population-screening. See you tomorrow.	general or unclear
3	"An invitation for population screening for cervical cancer and you call to make an appointment. Call back in 3 months because it's not being done now due to busyness in the practice. I have an opinion about that..."	cervical cancer
4	RIVM: every year dozens of deaths due to ignoring the smear test invitation... At half-past 12 on @NPORadio1 with department head Nynke van der Veen @rivm and @SPnl @HenkvGerven	cervical cancer
5	"I'll only be 30 in a week, but I already received the papers for the population screening. Tuesday a smear test brrrrrrrrrrr."	cervical cancer
6	BlvdD 12 March: had a mammogram for the first time, and it wasn't as bad as I expected. The woman doing it – I said, what a strange job this is – is happy about how many women she saves through early diagnoses. #population-screening	breast cancer
7	Invitation for population screening. Got my boobs squished again! :/	breast cancer
8	Oh great! You barely turn 50, and there's already an invitation for the boob-squisher on the doormat! #population-screening	breast cancer
9	At the GP's, the five-yearly female checkup #blech. Soon it will be over #population-screening	breast cancer
10	Just confronted with the fact that I'm almost 30. Magazine with an explanation about #population-screening on the doormat.	breast cancer
11	If we want to reduce mortality from colorectal cancer, and who doesn't, then screening/ #colorectal-cancer-research is the only realistic option.	colorectal cancer
12	Population screening for #colorectal cancer is starting gradually. Monday in #Ommelander Courant.	colorectal cancer

Table 27: Translated tweets with the correct classification of the supervised model for the category cancer type

	Tweet	Model	Original Dataset
1	What will change in cervical cancer screening from 2017 onwards? Visit our stand D073 #GPc2016 #GPconference	general or unclear	cervical cancer
2	"Quite annoying that yesterday I was informed by my GP that something was wrong with a mammogram from a year and a half ago and that I should have received a new invitation a year ago. #coronafailthanks Now Wednesday AvL."	cervical cancer	breast cancer
3	You participate in the cervical cancer screening. You get a bad result. Then you have to wait 6(!) weeks for further examination. There is so much wrong with our healthcare.	breast cancer	cervical cancer
4	Oh yes, population screening on the doormat.	breast cancer	general or unclear
5	Just had a discussion with ballet moms about mammography (which is worse:small breasts and then complaints that it doesn't show up, or large ones :that it doesn't fit	colorectal cancer	breast cancer

Table 28: Translated tweets with the incorrect classification of the supervised model for the category cancer type

	Tweet	Model	Original Dataset
1	Harmful consequence of national colorectal cancer screening: waiting times at gastroenterologists of 45+ days for people without screening but with symptoms!	Non-personal	Non-personal
2	I've been put in my place again. 1 letter for a smear test, 1 bill, and 1 letter from the tax authorities for the neighbours. That's great #valentine	Non-personal	Non-personal
3	good morning. It's so slippery. Called the GP, pap1, so it's good #smear test. So cervical cells were found #happy	Personal	Personal
4	Just been to the doctor for the population screening, done just like that. Hope the result comes by mail. Now coffee first #lovelylovely	Personal	Personal
5	Let women participate in a breast cancer screening programme from the age of 30. It would make a big difference; it would be discovered earlier.	Personal	Non-personal
6	in our programme @isala Astrid Nauta talks about the public academy on 14 March 2016, national colorectal cancer screening and new techniques.	Personal	Non-personal

Table 29: Translated tweets with the classification of the supervised model and the original labelled dataset for the category topic

	Tweet	Sentiment	Score
1	The result of my #population-screening #colorectaljancer is normal. Great! #gettested #stopcolorectalcancer	Positive	0.83
2	Great, turning 30. Now they want me to get a smear test.	Positive	0.86
3	Strong article in #Trouw ‘Screening colorectal cancer offers false promises’ by @caseofdees and @aliettejonkers #preventionparadox	Positive	0.53
4	On 25 March, it’s my turn again for @BVOZuidWest. #population-screening #breastcancer	Neutral	0.87
5	Damn, that hurt #population-screening	Negative	0.88
6	Today I got the result from the breast cancer population screening..... luckily nothing was found.	Negative	0.53
7	great’ birthday present! Invitation (!) for cervical cancer population screening #necessary #biteyourteeth	Negative	0.39

Table 30: Translated tweets with the sentiment analysis with certainty scores of the unsupervised model