Universiteit Leiden

Leiden Institute of Advanced Computer Science

Bachelor Informatica & Economie

Predicting sentiment on written medical self-assessments by elderly people using BERT and LIWC.

Bachelor Thesis

Rik van Baar

12/07/2022

Supervisors:

Prof.dr. M.R. Spruit

Dr. S. Verberne

ABSTRACT

As loneliness among elderly is increasing and the shortage of medical personnel is not decreasing, a method which keeps track of mental health could be used to reduce the workload and improve the healthcare received by elderly. Such a method is being developed in the Welzijn.AI programme in the form of an application. Part of this application involves predicting sentiment in short, colloquial, written Dutch medical texts by elderly. This thesis explored versions of both the word-based Linguistic Inquiry and Word Count (LIWC) model and the embedding-based Bidirectional Encoder Representations from Transformers (BERT) model to predict sentiment on these short texts. This was done in both a multiclass and binary setting. A dataset was acquired through a survey resulting in a collection of texts labelled as positive, negative or neutral. The objective was to see which of the models, or an ensemble of the two, would perform best and to evaluate the usability of the model in a real life application. To allow for a neutral class on the binary fine-tuned BERT model, a threshold analysis was done to determine the optimal thresholds for the neutral class. The experiments showed that using an ensemble of BERT and LIWC outperforms the separate models in both a binary (*F1: 0.98, kappa: 0.69*) and multiclass setting (*F1: 0.59, Kappa: 0.29*). It is concluded that future research is necessary to improve the multiclass model as it currently does not perform well enough to be used as a reliable model. When the question leaves no room for a neutral class the binary class can be used which performs well enough to use reliably.

Contents

1 Introduction

Research done by the Vrije Universiteit shows that more than a quarter of Dutch elderly (65+) are psychologically vulnerable (Campen et al., 2011). This vulnerability was measured through the amount of sad feelings, nervousness, anxiety, fearfulness and helplessness that the elderly experienced. The study shows that most people get vulnerable between the age of 65 and 85. A large psychologically vulnerable group of elderly can suffer from a series of disorders, including depression and anxiety. These problems arise both for the elderly and for the people in the inner circle, like family and friends, who might not even know about the seriousness of the situation. The Dutch government agency for economic forecasts and analysis 'Centraal Plan Bureau' also states that an increase in psychological help is necessary to reduce the amount of vulnerable elderly (Campen et al., 2011). The healthcare system however is characterized by a constant labour shortage and therefore the problem arises that there is not enough personnel available to monitor the psychological state of elderly on a regular basis. Even harder is monitoring the health of 'relatively healthy' elderly who live at home. Because they are healthy enough to live without daily care, keeping track of this group is hard as there are fewer contact moments than is the case with elderly in a nursing home for example. This is a problem because even though this group has fewer contact moments they might still experience serious mental issues. One of the points made by Rapport Taskforce De juiste zorg op de juiste plek (2018) in order to combat this inability was to embrace technological progress and use it to provide better and more efficient healthcare. This in combination with the advice by Van den Broek et al. (2021) that "Innovating in the area of IT can add to maintaining an active lifestyle as well as improving the quality of life" gives solid reason for a mobile health application that can monitor positive, neutral and negative sentiment over periods of time.

To provide in this need for an application and to help mitigate the problem of loneliness amongst elderly, the Welzijn.AI programme was initiated at Leiden University to investigate the possibilities of chatting in colloquial Dutch about their mental state to a mobile health application, after which a response is given through conversational AI. To be able to give a correct response, determining the sentiment of the text is essential. This thesis provides a model that can be used to predict sentiment for short, colloquial, written Dutch medical texts. Providing such a model can improve the performance of the mobile health application and provide insight to the attending physician or general practitioner. By monitoring the sentiment of the elderly over a longer period, physicians or other health workers are able to detect trends and provide the adequate care to those who need it the most. When used by elderly who live at home by themselves, there is also the added benefit that loneliness and sentiment can be easily detected, monitored and communicated to either family or an attending physician on a more regular basis without the patient feeling like they are subjected to a long series of questionnaires.

To predict the sentiment, the application receives text data which then has to be analysed, which puts this research in the natural language processing (NLP) domain. This subfield of AI focusses on the interaction between human language and computers. Detecting positive and negative sentiment in text is possible through a NLP technique called sentiment analysis which is one of the subfields of NLP that focusses on polarity detection as well as the detection of emotion towards something. Sentiment analysis is not new, and previous work has focussed on both traditional lexical methods like Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), and more modern machine learning techniques like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). LIWC is a supervised, word-based model that can be used to measure emotions by using the frequency of words in certain categories. This differs from a fine-tuned BERT model, which is an unsupervised and embedding-based machine learning approach that works with a self-attention layer (Jurafsky & Martin, 2019). The fine-tuned BERT model that is used also measures sentiment rather than emotion. As for the practical use cases, comparative research has been done between BERT and LIWC to rate emotion in psychotherapy (Tanana et al., 2021). RobBERT, which is a Dutch

version of BERT, and LIWC were also used to predict the presence of mental disorders based on psychiatric stories (Spruit et al., 2022).

This thesis will answer the question: To what extent can a combination of Bidirectional Encoder Representations from Transformers (BERT) and Linguistic Inquiry and Word Count (LIWC) accurately predict sentiment on written medical self-assessments by elderly people?
To research this, we look at which model is better at classifying these short Dutch medical texts, BERT or LIWC? (**Q1**) Since the only Dutch model that uses BERT sentiment analysis is a binary classification model, there is also the question of how we can add a neutral class to the BERT model and if this impacts the outcome of the predictions? (**Q2**) Additionally, the benefit when using the two models together to predict sentiment is researched. (**Q3**). Besides answering these questions, this thesis also tries to further understand the functioning of the model by analysing and comparing the errors of LIWC and BERT through an observational analysis and explainable AI techniques respectively. The last thing that will be discussed is how the results can be improved based on the findings of the error analysis. (**Q4**)

2 Background/Related work

2.1 NLP and Sentiment analysis

To understand the models that are used in this thesis it helps to understand the connection between natural language processing (NLP) and sentiment analysis (SA). Before this connection is explained, a brief overview of these concepts will be given, starting with NLP. The book "Natural language processing" by Kunar (2013) has the following description of NLP: "Natural language processing is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.". NLP is often seen as a subfield of AI and it tries to understand one or more of the 7 levels of language, which are: Phonology, morphology, lexis, syntax, semantic, speech, and pragmatic. Every level of language has its own obstacles. Ambiguity for example is a notorious problem within the semantic level as words can have different meaning in different contexts and therefore the context has to be understood by the computer. Generally, the more levels a program is able to understand the better. A full understanding of all levels could for example result in a fully automated dialog system (Khurana, et al., 2017). For this thesis the focus of the models is on the lexical and semantic levels as they are provide meaning to the word and meaning to the word in reference to the rest of the sentence respectively which is important to assess sentiment.

Sentiment analysis (SA) is one of the subfields of NLP and it focusses on polarity detection as well as the detection of emotion towards something. Sentiment analysis can be performed on the document level, sentence level and aspect level (Pandya, et al., 2020). SA on the document level tries to identify the sentiment of whole document while sentence level analysis focuses on parts of the sentence that might indicate a certain emotion or sentiment. There is also aspect-level analysis where certain aspects or topics are identified and analysed. This thesis will focus on document level understanding as people typically describe their experiences about something in multiple sentences. Sentiment analysis has been widely used to analyse social media posts and their corresponding sentiment while it also gained interest from the education domain (Kastrati et al., 2021). There are different techniques within NLP to assess sentiment, in general there are more traditional Lexicon based approaches and newer machine learning approaches that are used. The results of a challenge created by SamEval to predict sentiment in twitter messages also showed that ensembles can work well on sentiment analysis tasks, with half of the best submissions in the paper using an ensemble when achieving the highest score on both two-point and five-point ordinal scales (Rosenthal, et al., 2017).

2.2 Lexicon-based approach

Emotion analysis techniques based on lexical features have been around for quite some time. Work on modelling these connections with emotion analysis techniques increased around 2001 (Sadia et al., 2018). Since then, all kinds of techniques have been developed, of which LIWC is one of the most widely used (Lazarević et al., 2020). Linguistic Inquiry and Word Count is software that performs analysis on the frequency of words that occur in a text. Essentially, it counts how many times certain words are present in a sentence or a paragraph and reports back based on these values. The original LIWC model was made in 1993 with the updated revision being published in 2001. This updated revision used 74 output variables divided into 4 sub-categories (Standard linguistic dimensions, Psychological processes, Relativity and Personal concerns). Each of these sub categories contain more categories or output variables. An extensive list of all the categories can be found in the LIWC manual by Pennebaker et al. (2001). Throughout the years there have been improvements in the model with updates published in 2007 (Pennebaker et al., 2007) and 2015 (Pennebaker et al., 2015). With these updates came new categories while others disappeared, leading to a solidified model that has proven itself over the years. As compared to newer methods, there are disadvantages to consider when using LIWC, the biggest one being the method of classification. Because the nature of the model is simply counting words, improving the model based on a larger dataset is impossible. Only adjusting the dictionary or applying different pre-processing steps will lead to different results. Another disadvantage of this rule based methods is the multilingualism, different languages can have very different intonation if translated which can lead to misclassification.

In previous research, LIWC was used in the medical domain to evaluate emotion in Psychiatric stories (Spruit et al., 2022), rate emotion in psychotherapy (Tanana et al., 2021) as well as predicting depression status and severity (Liu et al., 2022). These researches concluded that, within their domain, LIWC is a useful tool to predict the respective health concerns. This research differs from other researches in the fact that it focusses solely on spoken text that is transcribed, as well as only targeting seniors and looking only at positive or negative sentiment instead of predicting symptoms like depression. In contrary to other previous research, Sun et al. (2020) concluded that: “LIWC positive and negative emotion dictionaries may not capture self-reported subjective emotion experience when applied to everyday speech ". Their research finds no significant correlation between the self-reported sentiment and the percentage of negative/positive words. This suggests that there is no use for LIWC in the model that is built in this thesis. However, this thesis focusses solely on the health domain where the questions specifically ask for the mental state of an elderly, therefore LIWC can still provide valuable results.

2.3 Machine learning-based approach

Because machine learning (ML) methods enable continuous improvements of a model, it is no surprise that machine learning advances are also made in the field of NLP. As compared to Lexicon based methods, ML learns from its experience. Within the NLP domain machine learning is often used to perform tasks like tokenization, part-of-speech tagging, named entity recognition and sentiment analysis. The advantage over traditional methods being that it is a scalable and trainable solution to develop a model. There are also no disagreements when categorizing words into groups as is the case with LIWC. These advantages of NLP resulted in the development of several methods. In this thesis, the Bidirectional Encoder Representations from Transformers (BERT) model is used. BERT is a model developed by Google that can be used to generate word embedding’s that are contextualized (Devlin et al., 2019). It does this by using transformers which map a sequence of input vectors to a sequence of output vectors of the same length while the Bidirectional in BERT means the input can be processed in the Self-Attention Layer from both left-to-right and right-to-left instead of casual left-to-right transformers (Jurafsky & Martin, 2019). This gives BERT the ability to understand the context of the text left and right from the word as it reads all words of a text in no specific direction. The BERT

framework consists of a pre-training and fine-tuning step (Devlin et al., 2019). When pre-training, the model is trained on a (very)large dataset of unlabelled data. For the fine-tuning step, the model is initialized with the pre-trained parameters, after which the parameters are fine-tuned with more specific labelled data (Devlin et al., 2019). This method of pre-training on a very large dataset and then fine-tuning on a smaller dataset with these already learned parameters is a form of transfer learning where knowledge gained form one model is used for a new problem. Transfer learning is usable because it allows to create powerful models with limited amounts of data if there is already a pre-trained model available. When used for classification, the BERT results are fed through a softmax function which normalizes the output of the model to a probability. The BERT model turned out to be useful when tested on a similar sentiment problem (Tanana et al., 2021), and the Dutch version RobBERT is open-source. When compared to a more traditional NLP-based ML method like the use of word features with tf-idf weights with support vector machines or naive bases, BERT also outperformed tf-idf on four different occasions while being significantly easier to implement (Gonz´alez-Carvajal & Garrido-Merch´An, 2021).

2.4 Dutch Versions

To ensure pervasiveness of linguistic accommodation, this thesis opted to use the Dutch version of BERT, called RobBERT (Delobelle et al. 2020) and a Dutch LIWC model rather than translating the words to English and running them through the English models. The publication by Boot et al. (2017) provided the most recent Dutch version of LIWC which essentially functions like the English model but with a Dutch wordlist. The first translation of the English wordlist was done in 2001 with some changes made in 2017 mostly adjusting it for the changes in categories. Results show that the English and Dutch LIWC models give equivalent results, with high correlation and low effect sizes (Boot, et al., 2017). This makes it possible to compare the results of the Dutch version to results obtained by the English version of LIWC.

The RobBERT model and associated paper by Delobelle et al. (2020) are used as the Dutch BERT model. RobBERT outperforms other Dutch BERT models in comparable research (Spruit et al., 2022). The model is based on the RobBERTa (Liu, et al., 2019) framework and trained on the Dutch section of the OSCAR corpus, which contains more than 126 million sentences. There is also a Dutch tokenizer instead of the default on which the model is pre-trained. The fine-tuning steps do in some cases have English tokenizers. Apart from this, the model is similar to RobBERTa and the underlying BERT model.

3 Methodology

3.1 Data acquisition

To validate the models, a labelled dataset was acquired through an online form. The form consisted of three fields where a text could be inputted. Participants were asked to formulate three different texts describing their week. One positive, a negative and a more neutral text. The form specifically asked to imagine oneself in a situation where they were a senior citizen (65+). Participants of all ages were allowed, and they were allowed to partake more than once. An impression of the form can be found in appendix 1. The main benefits of this method are that there is no need to label and that the dataset is guaranteed to be well-represented. Since there are always a positive, negative and neutral text submitted when the form is completed it is highly unlikely that an unbalanced dataset is obtained. To evaluate the texts, a quick scan is done to ensure the data quality. With this quick scan language and subject applicability are evaluated, only texts in other languages or texts that have nothing to do with the subject will be excluded from the dataset. The dataset contains 60 responses meaning there

are 180 labelled texts. Of these 180 texts none were removed during the quick scan, therefore a perfectly balanced dataset is obtained.

3.2 LIWC

As discussed in the background/related work section (2.4), one of the models that is used is the Dutch version of the LIWC model. The model uses different categories that are useful for this research. The most important being “PosEmo”, “PosFeel” and “Optimism” for the positive emotions and “Negemo”, “Anxiety”, “Anger” and “Sadness” for the negative emotions, with all categories weighing equally. A word can also occur in multiple categories, thus increasing the amount of negative or positive words in a sentence. For example, the sentence: “Ik heb net te horen gekregen dat ik meer medicijnen moet slikken voor mijn hoge bloeddruk dat is erg vervelend” has the negative word “vervelend” in both the “Anger” and “NegEmo” category. There is a possibility that this skews the result to the negative side, this is however the intended use of LIWC since a word is meant to be able to fall in more than one category. This is also partially nullified, since the same applies for the positive side. When using LIWC this way, there is no pure neutral class. Therefore, different thresholds will be explored to create a “neutral” class where the model is not confident enough to make a positive or negative prediction. For LIWC we just look at the difference in the amount of positive and negative words that occur in the text, with the threshold being the absolute difference between these two numbers. How the optimal threshold is found is explained in Multiclass classification (3.5). In most cases the words in the sentence undergo some form of stemming or lemmatization, as this was an off-the-shelf program however, there are no such steps in this Dutch LIWC model.

3.3 robBERT

The robBERT model is fine-tuned on the Dutch book review dataset, which contains more than 110k instances of labelled book reviews which are binary-classified as either positive or negative. The output of the BERT model is normalized to a certainty score through a softmax function that transforms the output to a probability. (Delobelle et al., 2020) Since this fine-tuned model is a binary classifier, a threshold has to be created for this model to allow a neutral class. Doing this is harder than it is with the LIWC model, since the output is normalized by the softmax function and therefore a probability threshold has to be determined. Determining the right threshold for BERT is explained in Multiclass classification (3.5).

3.4 Binary classification

To evaluate the different methods, it is interesting to look at how they perform on a binary classification task. Especially for BERT as it is trained on a dataset with binary labels and therefore should perform better without the addition of a neutral class. By performing a binary classification task, we can establish the optimal results and compare these to the multiclass classification results. Only the positive and negative answers are used from the dataset, meaning the neutral answers are ignored. This means that the results are best to be interpreted as the outcome if the question was more in the line of “Please explain why are you feeling good or bad today” This eliminates the possibility of a neutral class and therefore is more usable for a binary classification. Ignoring the neutral class also makes concluding based on the comparison between binary and multiclass classification harder. It is still useful however as it allows to compare binary to multiclass, especially since previous research found that binary sentiment classification will outperform multiclass classification when using LIWC (Salas-Zárate et al., 2017). The same is expected in this thesis since a more polarised dataset is used as compared to the multiclass so classifying should be easier. Because of the underlying application of the model, there is, besides comparison, relatively little use for this model within the scope of the thesis. It can however be used as a benchmark for further research, this means it is still useful to examine the results. The results will be evaluated using the F1-score, and

their inter-rater reliability will be assessed using Cohen's Kappa as this gives insight into the relationship between the two.

A problem with LIWC, as it is used in this thesis, is that it can have an equal amount of positive and negative words. When this happens, binary classification is impossible and the result will be labelled as "Undecided". This means that, apart from the F1-score, the amount of undefined results is another metric on which the model is assessed. The BERT model has no such issues, as it is highly unlikely that there is an equal probability because of the softmax function. When the Models are used together, the choice is made to classify a text as positive or negative as long as they don't oppose each other. For example, when LIWC classifies "Undecided" and BERT "Positive", the result is positive. When they do contradict, a "Undecided" label is given. This ensures that the model can be evaluated.

3.5 Multiclass classification

The most important results will come from the multiclass classification, as this is the method that is ultimately intended to be used in the mobile health application. To allow for multiclass classification, there are a couple of changes to be made to the models. Primarily, the addition of thresholds to define a new "neutral" class. Both models and their respective thresholds are discussed in the following section.

Adding a "Neutral" class to the LIWC model is similar to the "Undecided" class that was used in the binary classification, as it looks at the frequency of positive and negative words. The only difference being the range of the class which, instead of only classifying an equal amount of positive and negative words as "neutral", now ranges between 0 and 3 as the minimum absolute difference. Meaning that a text needs between 0 and 3 more positive or negative words to be classified as either positive or negative, depending on the threshold. Determining the optimal threshold is done through splitting the dataset into two random groups and testing all aforementioned thresholds. This is done 11 times, and the result is the average F1-score of these 11 iterations. The reason why this is done is to prevent overfitting and ensure stability of the model as much as possible. There are different techniques that are more common for determining these kinds of thresholds, like 10-fold cross validation. The problem is however that there are not nearly enough data points to use, and the model does not have to be trained. Therefore, this method is used as it still prevents at least a bit of overfitting. Apart from the F1-score, the LIWC model will also be evaluated based on the precision of the different classes. The most important one being the negative class, as an attending physician or general practitioner is more interested in potential negative signs, whereas a positive emotion that is misclassified as a neutral emotion is less severe.

Adjusting the BERT classifier and determining the right thresholds is done by looking at the probabilities which are outputted after the softmax function. Even though the softmax function normalizes the output of the model, it is still possible to distinguish lower probabilities from higher ones. As softmax normalizes the output between 0 and 1 the optimal threshold is believed to be between 0.96 and 1 judging from a small series of tests where a couple of example texts were classified. To find this optimal threshold, the same 11 times, 2-way, random split is used. The evaluation of BERT also favours negative precision over positive precision.

3.6 Ensemble model

BERT and LIWC will also be evaluated as an ensemble, which will lead to the answer of Q3. Combining the models will be done in 3 ways, the different methods are described in Table 1. Firstly, there is the model that does not have any favour towards BERT or LIWC, meaning that both predictions have the same weight. Within this option, there are also 2 differentiations. The first differentiation predicts "Neutral" when either one of the models predicts "Neutral" *(Model 1a)*, whereas the second differentiation predicts "Neutral" if and only if the two separate predictions contradict, and it predicts the non-neutral class otherwise *(Model 1b)*. The second option favours the LIWC model, this means that we only look at BERT if LIWC contradicts BERT, in this case a "Neutral" prediction is given *(Model*

2). In all other cases, the LIWC prediction will be leading. The third possibility is for the BERT model to be leading in the same way as described for LIWC. *(Model 3)*

Model 1a: ***Favours Neutral***

LIWC ↓, BERT →	**POS**	**NEU**	**NEG**
POS	POS	NEU	NEU
NEU	NEU	NEU	NEU
NEG	NEU	NEU	NEG

Model 2: ***Favours LIWC***

LIWC ↓, BERT →	**POS**	**NEU**	**NEG**
POS	POS	POS	NEU
NEU	NEU	NEU	NEU
NEG	NEU	NEG	NEG

Model 1b: ***Favours POS/NEG***

LIWC ↓, BERT →	**POS**	**NEU**	**NEG**
POS	POS	POS	NEU
NEU	POS	NEU	NEG
NEG	NEU	NEG	NEG

Model 3: ***Favours BERT***

LIWC ↓, BERT →	**POS**	**NEU**	**NEG**
POS	POS	NEU	NEU
NEU	POS	NEU	NEG
NEG	NEU	NEU	NEG

Table 1: Overview of classification rules

All models will be evaluated through the macro-averaged F1-score. This metric takes the mean of all the per-class F1-scores. As there is a perfect balance in sample sizes there is no difference between the Micro and macro F1-score. Just like with the separate models, the focus will be on the precision of the negative score to ensure as little serious misclassification as possible. The best performing model is used to analyse more closely, so an understanding of the errors can be examined, which will in turn lead to better understanding of the model and its potential strengths and weaknesses.

As an extra step, there will also be a comparison between the comparability of LIWC and BERT. Adding this evaluation metric provides insight into the chance adjusted comparability of the model. This tells us to what extent the models predict the same sentiment, this is useful for making statements about aspects of the model that are comparable or different. It can also be used in the error analysis. To execute this, both models will be run on the test set and the results of both contingency tables will be added. After this the comparability can be assessed through Cohen's kappa. Kappa will tell nothing about the actual performance of the model, since there could be a case where both LIWC and BERT predicted the same, but also the wrong sentiment. Therefore, we analyse both Cohen's Kappa and the F1-score to give a more complete picture of the model.

3.7 Error analysis

To understand the model an error analysis will be performed. Such an analysis can be done in multiple ways. Often, research on machine learning methods have an ablation study where one or more of the input features are left out to assess the impact of certain features on the model. Since the Dutch robBERT model only takes the text as an input and is pre-trained there are no such possibilities where a feature can be cross validated against other features. Methods which do work for this kind of textual input are SHAP (Lundberg, et al, 2017) and LIME (Ribeiro, et al, 2016), these both visualize the importance of features to the classification result. Another well-known and usable method is BERTviz (Vig, 2019). BERTViz visualizes the attention between the model layers in the BERT model to give insight into its behaviour. This attention in the model is given to the tokens which are determined by the tokenizer. By visualizing this attention, it is possible to say to which extent a token has influence on the final prediction.

As LIWC is a word-based model it is a lot more analysable as the output is simply the amount of words that fall in the positive or negative categories that were created. Explaining the model can be done by looking at the words that are classified as positive or negative with and without the context of the sentence, if there is a substantial difference in meaning then this tells something about the shortcomings of the model. The model will be examined by taking some edge cases where there are almost the same amount of positive and negative words as well as some misclassified texts. This is done to get insight into the strengths and possible shortcomings of the model.

4 Results

This section contains all the results of the LIWC and BERT models and their collective performance, the section has the same order as the methodology section as it starts with the results of the binary classification (4.1). Subchapter 4.2 describes the threshold tuning for the multiclass classification, after which the comparability is tested (4.3) and the different models are evaluated (4.4). Finally, there is the error analysis (4.5).

4.1 Binary classification

Table 2 shows the results of the binary classification where *n* is the amount of classified results as compared to the total instances. A result can, in this case, only be classified as "Undecided" when there is an equal amount of positive and negative words in a text or if the BERT model predicts exactly 0.5. All cases of "Undecided" are from the LIWC model however, since the BERT model almost never predicts a chance of 0.5 since it uses a softmax activation function that normalizes the result. The total sample size is 120 since only the positive and negatively labelled data were used, this excludes the neutral labelled data. This is done to allow the binary classification, this also means that unclassified results are either positive or negative and not neutral.

LIWC

Pred ↓, Act →	**POS**	**NEG**
POS	49	10
NEG	3	31

n = 93/120

Precision	0,831
Recall	0,942
F1	**0,883**

BERT

Pred ↓, Act →	**POS**	**NEG**
POS	57	14
NEG	3	46

n = 120/120

Precision	0,803
Recall	0,950
F1	**0,870**

BERT + LIWC

Pred ↓, Act →	**POS**	**NEG**
POS	48	1
NEG	1	24

n = 74/120

Precision	0,980
Recall	0,980
F1	**0,980**

Table 2: Binary classification results for LIWC, BERT and the combined model

It is clear that the binary classifier performs very well. The model is good at predicting whether a text is positive or negative when provided with a polarized dataset, the best performing model being the combination of BERT and LIWC. When looking at how many of the instances are actually classified however, we see that this increase in accuracy comes with a decrease in classified results, with the combined model labelling 46 results as "Undecided". When only looking at BERT we see a F1-score of .87 which is a good score that is usable for real world applications. The downside of this binary classification still being that it is not suiting the use-case of this thesis, which is to do a multi class classification. The binary classification does however provide a context for the multiclass model to operate in.

The evaluation of the comparability is done through Cohen's Kappa, which determines the comparability of a result through a chance-adjusted metric that shows how much better a model performs than if it were to be a random guess. The result is a number between 0 and 1, with 1 being a perfect model and 0 being a model that performs the same as a random guess. The result of 0,69 means the comparability of the models have 'substantial agreement' judging from the model by Landis & Koch (1977). The results can be found in Table 3.

Intercomparability		**n = 62/120**
LIWC ↓, BERT →	**POS**	**NEG**
POS	34	2
NEG	7	19

Pe = 0,526
Po = 0,855
Cohen's Kappa = 0,694

Table 3: Comparability of LIWC and BERT, binary

4.2 Threshold tuning

When looking at the multiclass classification, the first results are from the threshold tuning. This is done both for BERT and LIWC, with the results for BERT in table 4 and for LIWC in Table 5. Both models show the iterations against the threshold. The BERT models shows a comparable result for the different thresholds with the best threshold being *0.995*. This result means that, over 10 iterations of a split evaluation set, the best suited threshold is a probability level of *0.995*, below this probability the model should predict "Neutral" instead of "Positive" or "Negative". The reason why this number is so close to 1 is because of the softmax function which is used by BERT to normalize the output of the model. The reason why the thresholds lower than *0.995* result in worse average F1-scores is due to the fact that there is a more skewed class distribution since more instances will be classified as "neutral" causing the F1-score to drop.

The thresholds for LIWC show a similar pattern with a declining F1-score when the threshold allows for more "Neutral" classifications. The threshold is the minimal absolute difference between the amount of positive and negative words necessary to predict either of these classes. In this case no threshold at all turns out to work the best on the sentiment data. The reason for this is that there is, even without a threshold, a large neutral class as there are quite some instances where the positive words equal the negative words. In this case the model predicts the text as neutral. Increasing the threshold only leads to more neutral predictions and, as was the same with BERT, to a more skewed class distribution.

Threshold tuning, BERT

Threshold → *Iteration ↓*	**0,97**	**0,98**	**0,99**	**0,995**	**0,999**
1	0,542	0,527	0,543	0,602	0,583
2	0,543	0,535	0,547	0,583	0,602
3	0,543	0,551	0,534	0,575	0,615
4	0,582	0,588	0,595	0,589	0,625
5	0,614	0,598	0,609	0,631	0,611
6	0,572	0,570	0,567	0,600	0,595
7	0,566	0,554	0,555	0,597	0,586
8	0,581	0,573	0,586	0,618	0,588
9	0,552	0,554	0,568	0,613	0,601
10	0,583	0,572	0,581	0,603	0,564
11	0,596	0,596	0,605	0,635	0,535
Average	0,570	0,565	0,572	**0,604**	0,591

Table 4: BERT threshold tuning results

Threshold tuning, LIWC

Threshold → *Iteration ↓*	**0**	**1**	**2**	**3**
1	0,611	0,526	0,449	0,309
2	0,642	0,611	0,498	0,365
3	0,550	0,476	0,313	0,229
4	0,551	0,549	0,450	0,340
5	0,672	0,613	0,521	0,392
6	0,575	0,534	0,447	0,375
7	0,655	0,612	0,433	0,340
8	0,645	0,615	0,493	0,334
9	0,594	0,582	0,461	0,289
10	0,552	0,541	0,424	0,308
11	0,655	0,613	0,463	0,358
Average	**0,609**	0,570	0,450	0,331

Table 5: LIWC threshold tuning results

4.3 Comparability

Table 6 shows the comparability of both models on a multiclass scale. As can be seen, Kappa is 0.26 which is a 'fair' result according to Landis & Koch (1977). The addition of the third class impacts the score by 0.43 as the score for binary classification was 0.69. The low Kappa score however only tells how good the models are at predicting the exact same. The model says nothing about the actual performance of the models, so this metric does not immediately suggest a 'bad' model. When looking at the table most misclassified data has to do with the "Neutral" class and there are only 2 cases of a real contradiction between positive and negative. The data from this table will be used to analyse the models and to highlight some differences in the error analysis part.

Intercomparability

LIWC ↓, BERT →	**POS**	**NEU**	**NEG**
POS	44	7	1
NEU	15	49	38
NEG	1	4	21

Pe = 0,344
Po = 0,512
Cohen's Kappa = 0,263

Table 6: Comparability of LIWC and BERT, multiclass

4.4 Model Evaluation

This chapter discusses the different models and their performance on multiclass sentiment analysis. As discussed in Multiclass classification there are 3 models that are used. All models are evaluated on their average macro F1-score and compared against each other. Since there is an emphasis on negative emotions, as they are the most important to find, we also look at the recall and precision of the model for the "NEG" class. Model 1a has 3 tables instead of 1, the extra tables contain the separate LIWC and BERT scores, so they can be compared to the other models. All other tables show the combined model, as there is no difference in the single BERT or LIWC evaluation

Model 1a: LIWC

Pred ↓, Act →	POS	NEU	NEG
POS	49	20	10
NEU	8	31	19
NEG	3	9	31

	Precision	Recall	Macro - F1
POS	0,620	0,817	0,705
NEU	0,534	0,517	0,525
NEG	0,721	0,517	0,602
Average	0,625	0,617	**0,611**

Model 1a: BERT

Pred ↓, Act →	POS	NEU	NEG
POS	53	18	8
NEU	6	16	15
NEG	1	26	37

	Precision	Recall	Macro - F1
POS	0,671	0,883	0,763
NEU	0,432	0,267	0,330
NEG	0,578	0,617	0,597
Average	0,560	0,589	**0,563**

Model 1a: BERT + LIWC

Pred ↓, Act →	POS	NEU	NEG
POS	44	7	1
NEU	15	49	38
NEG	1	4	21

	Precision	Recall	Macro - F1
POS	0,846	0,733	0,786
NEU	0,480	0,817	0,605
NEG	0,808	0,350	0,488
Average	0,711	0,633	**0,626**

Table 7: Model 1a - LIWC, BERT and Combined models with the focus on neutral classification

Starting with model 1a we look at Table 7, these tables describe the performance of LIWC, BERT and a combination of the two respectively. Model 1a focusses on precision over recall as it classifies cases as “Neutral” if either one of the models predicts “Neutral”.

When looking at the LIWC model there i a relatively high “NEG” precision as well as a high recall for “POS”. There is also a substantial gap between the positive, neutral and negative F1-scores. This high Recall for the positive class suggests that the LIWC model captures most of the positive cases in the dataset, the precision is substantially lower however, meaning there are still quite some false positives. A stronger result on the positive class was expected and is comparable to research by Filho et al. (2013) where they concluded that LIWC performs better on indicating positivity than negativity. When averaging the result to a macro F1-score we find that it is comparable to the threshold score of 0.609.

The BERT model has high precision and recall for the positive class suggesting that it is good at identifying positive instances. The Neutral class however sees both low precision and recall. The low recall could indicate that the threshold is too high as there are not enough neutral cases identified. If we check this with the contingency table this seems to hold up as there are only 37 neutral cases against 73 and 79 negative and positive cases respectively. Adjusting the threshold however would not necessarily positively impact the precision or the F1-score.

A combination of LIWC and BERT shows a better result than the other two models, suggesting that combining these models improves performance. Since the model favours neutral, the high recall and low precision for the neutral class is explainable as a larger share of the instances is classified neutral. Noticeable is the low recall of the negative class, when looking at the table it can be seen that there are a lot of negatives classified as Neutral which explains this score. The combined model performs marginally better on positive classification as well.

Model 1b: BERT + LIWC

Pred ↓, Act →	POS	NEU	NEG
POS	56	22	1
NEU	3	16	22
NEG	1	16	36

	Precision	Recall	Macro - F1
POS	0,709	0,933	0,806
NEU	0,390	0,296	0,337
NEG	0,679	0,610	0,643
Average	0,593	0,613	**0,595**

Table 8: Model 1b - ensemble with the focus on positive or negative classification

Model 1b, which is shown in table 8, favours positive and negative classifications over neutral classifications. For this model a high F1-score can be found for the positive class, which is comparable to 1a. The difference is found in the neutral and negative class with the neutral class having a lower F1-score and the Negative class having a higher F1-score. The lower neutral score is due to the fact that there are only 41 neutral classifications as compared to the 102 of 1a. This decreases the recall and therefore the F1-score. This decrease in neutral classifications does help the negative recall as a part of the neutral classifications of model 1a are now classified as negative. The downside being that the precision decreases by 0.1.

Model 2: BERT + LIWC

Pred ↓, Act →	POS	NEU	NEG
POS	49	13	3
NEU	10	40	30
NEG	1	7	27

	Precision	Recall	Macro - F1
POS	0,754	0,817	0,784
NEU	0,500	0,667	0,571
NEG	0,771	0,450	0,568
Average	0,675	0,644	**0,641**

Table 9: Model 2 – ensemble which favours LIWC over BERT

Table 9 describes the results of model 2. This model favours LIWC in the way that only if the models contradict on a positive and negative classification the result is neutral. All other cases favour the prediction by LIWC. This model again shows a high positive F1-score with lower values for the neutral and negative classes. The negative precision is higher than model 1b and comparable with the combined model of 1a. There is a drop in recall when compared to 1b which causes the F1-score for negativity to be lower than 1b.

Model 3: BERT + LIWC

Pred ↓, Act →	POS	NEU	NEG
POS	51	16	4
NEU	8	25	26
NEG	1	19	30

	Precision	Recall	Macro - F1
POS	0,718	0,850	0,779
NEU	0,424	0,417	0,420
NEG	0,600	0,500	0,545
Average	0,581	0,589	**0,581**

Table 10: Model 3 - ensemble which favours BERT over LIWC

The last model that was evaluated is model 3, this model favours BERT over LIWC using the same logic as model 2. As table 10 shows, this model does not perform better than the other models. Since this model favours BERT and model 1a and 1b showed a substantial difference in the neutral F1-score it is within expectations that, albeit less substantial, BERT has a lower neutral-F1-score. The reason being that BERT predicts less Neutral classes than LIWC. This makes the recall differ 0.2 between model 2 and 3. This lower recall adds some examples to the positive and negative class causing them to rise.

To make comparison easier, table 11 shows all average precision, recall and macro F1-scores from the different models.

Model	Precision	Recall	Macro - F1
1a: LIWC	0,625	0,617	**0,611**
1a: BERT	0,560	0,589	**0,563**
1a: BERT + LIWC	0,711	0,633	**0,626**
1b: BERT + LIWC	0,593	0,613	**0,595**
2: BERT + LIWC	0,675	0,644	**0,641**
3: BERT + LIWC	0,581	0,589	**0,581**

Table 11: overview of Precision, Recall and F1-score of different models

4.5 Error analysis

Table 12 shows 5 examples of misclassified texts and the corresponding output which we will evaluate. The table shows some common issues with LIWC with the majority of the errors occurring because LIWC was not designed to interpret context. The errors that are found are: a positive word that is classified as negative or the other way around, an adjective which changes the meaning of the classified word is not taken into consideration, ambiguity and cases where words are missed because they are not in the right stemmed form.

Nr	Sentence	Sentiment Predicted	Sentiment real	Predicted words
1	*Niks is goed, laat me met rust!*	*positve*	*negative*	*PosEmo: [('goed', 1), ('rust', 1)]*
2	*Stevig inkomensverlies sinds pensionering beperkt m'n mogelijkheden wel meer dan mij lief is.*	*positve*	*negative*	*PosEmo: [('lief', 1)]* *PosFeel: [('lief', 1)]*
3	*Die rot knie ook, hoop dat die prik nu beter helpt*	*positve*	*negative*	*PosEmo: [('hoop', 1), ('beter', 1)]* *optimism: [('hoop', 1)]*
4	*Nou ik heb zo'n leuke dag gehad met Thea. Helemaal geen last ook van de operatie van laatst!*	*negative*	*positive*	*NegEmo: [('last', 1)]* *anger: [('last', 1)]*
5	*De hele week super slecht weer, niks kunnen doen dan alleen maar binnen zitten, gelukkig wordt het morgen iets beter, dan zou ik eindelijk een rondje kunnen lopen*	*positive*	*Undecided*	*PosEmo: [('super', 1), ('gelukkig', 1), ('beter', 1)]* *PosFeel: [('gelukkig', 1)]* *optimism: [('super', 1)]* *NegEmo: [('slecht', 1), ('alleen', 1)]* *sadness: [('alleen', 1)]*

Table 12: LIWC errors

When looking at sentence 5 we see “super slecht” which translates to “super bad”. “Super” in this case is labelled as positive as it has a positive meaning when used by itself. In this case however, “Super” is an adjective to “slecht” which means bad and therefore it should be labelled as negative as it is used as an emphasis for the negative word. This inability of recognizing certain simple adjectives

is also seen when looking at sentence 4. It is common to have adjectives change the meaning of a word. For example, “geen last” roughly translates to “no pain”. LIWC however only sees the word “last” which means pain, and classifies it as a negative emotion. This completely ignores the actual positive meaning of the sentence that describes how there is now no more pain. The same goes for the first sentence, “Niks is goed” meaning “Nothing is going well”. LIWC only recognizes “Well” and does not understand the connection between “Nothing” and “Well”.

There is also the case of ambiguity where words have double meanings which are easy to get wrong, an example being “lief” in sentence 2. In the sentence “lief” is used to describe something negative, even though the word itself often has a positive meaning. This again is only possible to distinguish when looking at the context of the word in a sentence which is impossible for LIWC. To give another example of the double meanings, the word “alleen” can be translated to “Alone”, “Lonely” and “Only” in English. In sentence 5 it is meant as “Only” which can’t be really classified as a positive or negative emotion, but because “alleen” has different meanings in Dutch it classifies as negative as it is seen as “lonely”.

Another thing that is common is the case where there is a conjugation of a word present in the text which doesn’t get recognized because only the stemmed version of the word exists in the LIWC library. This happens in sentence 4 where “leuke” is not classified because it a conjugation of “leuk” which means nice. “Leuk” however is in the Positive emotion category.

To analyse the results of BERT the intended method was to use SHAP, LIME or BERTViz as it visualises the model output which is useful when determining what causes a wrong classification. The interpretability of the results coming from these tools are less interpretable then initially thought however. This is caused by the tokenizer which is used during the fine-tuning on the robBERT-dutch-books model. This tokenizer splits the text into English words and therefore it is impossible to see which Dutch words are causing a certain prediction through attention visualisation as Dutch words do not provide meaning as whole words but rather as smaller strings of English words. There is a v2 version of the tokenizer which is Dutch, however the fine-tuned model is trained on the old English tokenizer. As the model still performs better than random there is meaning in the different tokens even though they are not Dutch words. This analysis would however take considerably more time as interpretation is more difficult and therefore it is deemed beyond the scope of the research.

5 Discussion

The question to what extent a combination of the Bidirectional Encoder Representations from Transformers (BERT) machine learning technique and the Linguistic Inquiry and Word Count (LIWC) technique can accurately predict sentiment on written medical self-assessments by elderly people will be answered through the 4 sub-questions defined in the introduction. During the research there were limitations that might influence the outcome of the research. These will be addressed in the limitations section (5.1), after which the sub-questions will be answered (5.2).

5.1 Limitations

The limitations of the research are best described when divided into the categories Data, Thresholds, Models and Error analysis.

This research only had 180 instances divided into three classes. Although this was enough data to work with, as the BERT model was pre-trained and therefore did not rely on large amounts of training data, the results would be more accurate if a larger dataset had been used. Because of the small amount of instances, it was also not possible to improve the BERT model through fine-tuning. For future research it would be a good starting point to acquire more data and fine-tune the BERT model as it would most likely yield better results.

Threshold-tuning as it is done in this research is unconventional and it is only the result of several limitations of the data and models. Ideally there would be no threshold tuning as the BERT model would define those boundaries when it is trained. However, due to the absence of enough training data this was not possible and an artificial neutral class was introduced. The threshold-tuning was done as random as possible with the average score of ten random macro F1-scores used per analysed threshold. More conventional methods like 10-fold cross validation were also not possible as there is no training and evaluation split to be made.

The models also had limitations as the intended use is different from how they were used in this research. This resulted in some limitations which lead to concessions that had to be made. Most notably adjusting the LIWC model to predict sentiment instead of emotion, and creating a neutral class in the binary BERT classifier. The LIWC model is intended to give a percentage of words that fall within a category, not predicting if a text is positive or negative. The solution to this was using the categories for positive and negative emotions as well as some other emotions to create the artificial positive and negative categories. The concession here is that it is not extensively researched which categories should or should not be included in the artificial classes and therefore possibly limiting the outcome of the model. The pre-trained RobBERT model is a binary classifier and introducing a neutral class after training is a limitation of the research. Ideally there would be enough labelled data to fine-tune the model on a multiclass dataset as the model then actually learns to predict neutral classes instead of classifying a texts as neutral when the probability is not high enough to be classified as either positive or negative.

The error analysis also has some limitations, especially with BERT. Because of the English tokenizer it was not possible to usefully analyse the attention through BertVIZ within the scope of the research. Other methods like SHAP and LIME are also not able to give the insight that is neededto properly analyse the decisions. The attention is still interesting because the classification results are better than random guesses, but clearly following the models attention through Dutch words is impossible with this fine-tuned model which limits interpretability.

5.2 Sub-questions

(**Q1**) **How do BERT and LIWC compare to each other when predicting sentiment on written health self-assessments by elderly people?**

To answer this question a distinction has to be made between the binary and multiclass predictions. The results of the binary classification show a very similar F1-score and therefore the case can be made that they perform almost identical on this classification task. There is however a substantial nuance to be made on this statement as it does not include negative scores for the “Undecided” class the LIWC model outputs. The LIWC model only classifies 93 of 120 instances where the BERT model answered all. This difference comes from the instances where LIWC finds the same amount of positive and negative words. As BERT is trained on binary classification and the output is normalized through the softmax activation function no such problems arise for this model. Both models do perform well on the binary classifications with F1-scores of 0.88 for LIWC and 0.87 for BERT. When looking at the recall and precision, as well as the contingency table, no large differences are found besides the number of classified instances. Because of the difference in instances classified and the corresponding F1-scores of both models it is concluded that BERT is better suited for predicting sentiment on self-assessments if it is a binary classification problem.

For the multi class classification a neutral class was added which the thresholds were artificially determined through a series of tests, Q2 discusses this neutral class to more extent. This new neutral class eliminated the possibility for LIWC to predict “Undecided”, instead an equal amount of positive and negative words became the “Neutral” class. The result of this new class shows

a higher score of 0.61 for LIWC as compared to the 0.56 for BERT. Especially the F1-score for the neutral class is big, this suggests that, for the thresholds chosen, BERT is significantly worse at predicting if a text is neutral. This is not a surprise as the model is pre-trained on a large set of book reviews which were binary labelled. Both the BERT and LIWC model are better at predicting positives than negatives which for LIWC is also in line with the research of Filho, et al. (2013) who concluded the same. Apart from these differences the models perform similarly, therefore LIWC is better suited to use when predicting sentiment on multiclass classification within the scope of this thesis. The comparability of the multi-class model was also assessed through Cohen's Kappa as this provided an insight into the agreeableness of the model. With a kappa of 0.26 the models are fairly comparable, this is mostly due to the neutral class as can be seen in the contingency tables in the results section.

In conclusion, even though for the binary classification task the F1-scores are very similar, the BERT model outperforms the LIWC model because it classifies all instances instead of only 94 of 120. For multiclass classification the LIWC model scores higher on all important metrics and is therefore better suited for the classification task.

(**Q2**) **How can a neutral class be added to the BERT model and does this impact the outcome of the predictions?**

To answer this question, we will look at the thresholds used to define this neutral class and then discussing the change in performance. Since both the Dutch LIWC and pre-trained robBERT model are not designed for multiclass classification there was a challenge defining this neutral class.

Creating this neutral class was done with a threshold analysis in which different thresholds were tested against each other. This was done through 11 iterations for every threshold, where each iteration was half of the total dataset which was randomly chosen without doubles. This way, a new class was created. As for the performance, the results showed that for LIWC it was best to only classify as neutral if there are an equal amount of positive and negative words. For BERT the best found threshold was a 0.995 probability. All predictions below this are classified as neutral. The F1-score for LIWC was 0.27 lower than it was without the neutral class. For BERT this difference was 0.32, it is important to know the dataset used to evaluate the model without a neutral class also did not contain neutral cases. This difference is quite substantial, and it is very clear that both models perform better on binary problems than multiclass. This is in line with other research done about LIWC (Salas-Zárate et al., 2017). Overall, if possible binary classification is preferable over multiclass classification on this specific task. The scope of the thesis however does not allow this, therefore multiclass classification is still used.

(Q3) Is there is a significant benefit when using an ensemble of LIWC an BERT to predict sentiment?

With the goal of the research being a model which accurately predicts sentiment a combined model of LIWC and BERT would only solidify the results and is therefore very interesting. As becomes clear from the results both the ensemble and the separate models have their own benefits. The binary ensemble showed a F1-score of .98, thereby outperforming all other models that were tested. With a Kappa of .69 there is also "Substantial agreement" which in combination with the high F1-score means this is a strong ensemble. However, the amount of undecided classifications is substantially higher than with the separate models which is a downside because a lot more manual classification is necessary. The separate BERT model did classify all instances, but had a lower F1-score and is therefore less usable to accurately predict sentiment. LIWC is not preferred as there is no substantial increase in F1-score while the number of undecided classifications is quite large. When looking at the results it is concluded that the ensemble should be used when looking for the most accurate result possible. When the goal is to obtain the highest coverage, the BERT model should be used. If we take the scope of the research into consideration, the ensemble is the most useful as it provides a very

strong model which can be accurately used to predict sentiment. This is important as wrong predictions might be harmful and should be reduced to a minimum.

The ensemble of the multiclass classification has 4 different configurations as described in the methodology section. We tested model 1a, 1b, 2 and 3. As LWIC scored 0.61 on the solo test the combined models should achieve a higher F1. This happened for model 1a and 2 with scores of 0.64 and 0.62 respectively. This is a marginal increase in score and a further look into the F1-score per class, precision and recall gives us a bit more insight into which model performs best. On average the Positive F1-score is still very strong with the lowest being the LIWC F1-score of 1a. Model 1b has the highest positive F1-score and should be used when the focus is on the classification of positive results. Predicting the neutral class is the hardest across the models with only model 1a managing a 0.6 F1-score. As we favour a strong negative class F1-score a selection based on that should also be made. When we look at it only model 1b has a higher negative F1-score than the single LIWC model. However, 1b does lose out on the neutral F1-score as especially its recall is lower than the other models. If we then look at the model which best fits the needs for the mobile health application 1b is the model with the highest positive and negative F1-score, therefore it is better suited for the classification task of this thesis. This is despite the fact that model 2 has a higher average F1-score and theoretically performs better. It is shown that ensembles can increase the result in both a binary and multiclass setting and they are the preferable model in both classification tasks, therefore it can be concluded that there is a significant benefit when using an ensemble to predict sentiment.

(Q4) How could the results be improved based on the findings of the error analysis.

The error analysis showed the errors made by LIWC and gave a possible explanation why these occur. Improvements could be made by either improving the model, the pre-processing or the artificial categories. Improving the model itself is only possible through a more accurate set of categories or a more representative wordlist. This is hard as it requires a lot of additional research and it might be the best to wait for a new version by Boot, et al. Another option which is easier to implement is to adjust the range of categories that is taken for the positive or negative class. Currently “PosEmo”, “PosFeel” and “Optimism” are used for the positive emotions and “Negemo”, “Anxiety”, “Anger” and “Sadness” for the negative emotions, excluding or including some categories could improve the model. The third method of improving the model as used in this research is by applying pre-processing steps like unit-normalisation, stop words removal or lemmatization. Other measures like unit standardisation or number standardisation might improve the results of LIWC.

Improving the results form BERT is less straightforward as it requires extensive research into the decisions made by the model. This is hard because of the current tokenizer which is based on the English corpus which causes techniques like BERTViz, SHAP and LIME to be less effective for interpretation then with a Dutch Tokenizer. Fine-tuning with a Dutch tokenizer is a way to make interpreting easier because the attention of the model on full Dutch words can be visualized. Even though it might not positively affect the results it can be argued that a more interpretable model is a better model if other conditions remain constant. Other improvements are possible if better pre-trained models, tokenizers or fine-tuned versions become available.

6 Conclusion

This research aimed to find a model or ensemble that accurately predicts sentiment on Dutch written medical self-assessments by elderly people, which is done to address the issue of loneliness among elderly in the Netherlands. The model will be part of a future application that allows elderly to chat in colloquial Dutch about their mental state to a mobile health application. This gives general practitioners or others involved with the health of an elderly the option to keep track of the sentiment over a period of time, allowing them to act on negative signals.

A BERT and LIWC model were used to classify short self-assessments that were acquired through a survey. Both models were altered so they could perform multiclass classification and predict a positive, negative and neutral class. Adjusting the models was done by introducing artificial positive and negative classes to LIWC and defining a threshold based on the probability score for BERT.

Based on the results of the two models, it can be concluded that using an ensemble of BERT and LIWC that favours positive and negative classifications over neutral classifications is best used when predicting sentiment on Dutch written medical self-assessments by elderly people. Even though this ensemble has a macro F1-score of 0.59, which is lower than the highest found score of 0.64, the ensemble scores higher on the F1-scores of both the positive and negative class which is deemed more important than the neutral F-1 score. The ensemble also outperformed the separate models on most occasions proving the usability of a combination of the two models, especially when LIWC precedes BERT in the decision hierarchy as is the case in model 2. When choosing only one model, LIWC outperforms BERT in both the binary and multiclass classification task, therefore making LIWC better suited when only choosing one model.

When removing the neutral class and making the classification binary, it is shown that both models are better at predicting the right sentiment than in a multiclass setting. With the ensemble as a binary classifier, where a neutral prediction is a separate “Undecided” class, a F-1 score of 0.98 is realized. This is better than the separate scores for LIWC and BERT and therefore it is concluded that for binary classification on Dutch written medical self-assessments an ensemble outperforms both the BERT and LIWC models as used in this thesis.

Error analysis showed that LIWC mainly has problems with ambiguity as it misinterprets word because of the lack of context. Contextualising is the hardest part for LIWC as it is merely a word-based model and therefore it ignores, sometimes critical, context which might change the meaning of words. BERT is able to contextualise, but analysing BERT was beyond the scope because of the English tokenizer that is used for fin-tuning.

Future research should focus on better understanding the decisions made by BERT so it can be improved. It is also advised to increase the dataset and use this to fine-tune BERT as a multiclass classification model rather than a binary model. This eliminates the need for the fine-tuning step and allows for a more accepted validation approach like 10-fold cross validation. Another thing which can be researched is the effect of including or excluding categories in the artificial sentiment classes of the LWIC model as well as adding pre-processing steps to the LIWC model.

Finally, the multiclass ensemble currently does not perform well enough to be used directly as a method of keeping track of the sentiment of elderly through an application. It can however be used as an advisory tool which is used in parallel with the current processes rather than functioning as a substitute. When rephrasing the question asked by the application to only allow for a positive or negative answer the binary ensemble is usable to accurately predict sentiment and can give valuable insight to general practitioners, family members or others involved on the wellbeing of elderly.

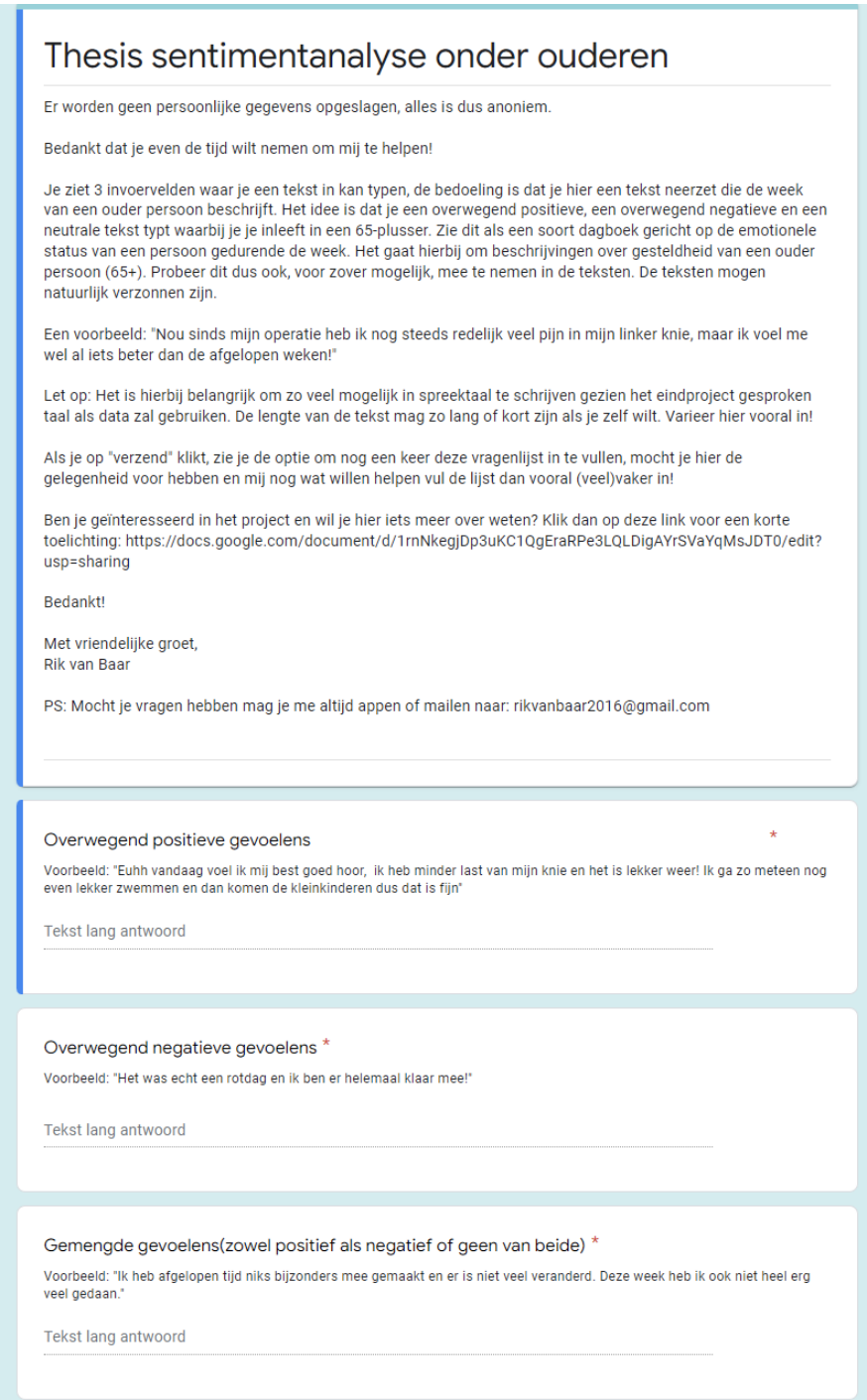
7 References

1. Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. Dutch Journal of Applied Linguistics, 6(1), 65–76. https://doi.org/10.1075/dujal.6.1.04boo
2. Campen, C., Ross, J. A., & Van Campen, C. (2011). Kwetsbare ouderen. Sociaal en Cultureel Planbureau. http://docplayer.nl/2498379-Kwetsbare-ouderen-cretien-van-campen-red.html
3. Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. Findings of the Association for Computational Linguistics: EMNLP 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.292
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1423
5. Filho, P.P., Pardo, T.A., & Aluísio, S.M. (2013). An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. STIL.
6. Gonz´alez-Carvajal, S., & Garrido-Merch´An, E. C. (2021, januari). Comparing BERT against traditional machine learning text classification. https://doi.org/10.48550/arXiv.2005.13012
7. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing. Els autors. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf
8. Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. Applied Sciences, 11(9), 3986. https://doi.org/10.3390/app11093986
9. Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges.
10. Kumar, E. (2013). Natural Language Processing. Penguin Random House. https://books.google.gg/books?id=FpUBFNFuKWgC&printsec=copyright#v=onepage&q&f=false
11. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159. https://doi.org/10.2307/2529310
12. Lazarević, L. B., Bjekić, J., Žlvanović, M., & Knežević, G. (2020). Ambulatory assessment of language use: Evidence on the temporal stability of Electronically Activated Recorder and stream of consciousness data. Behavior Research Methods, 52(5), 1817–1835. https://doi.org/10.3758/s13428-020-01361-z
13. Liu, T., Meyerhoff, J., Eichstaedt, J. C., Karr, C. J., Kaiser, S. M., Kording, K. P., Mohr, D. C., & Ungar, L. H. (2022). The relationship between text message sentiment and self-reported depression. Journal of Affective Disorders, 302, 7–14. https://doi.org/10.1016/j.jad.2021.12.048
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
15. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
16. Ministerie van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer. (2022, 17 mei). Aanpak eenzaamheid onder ouderen. Eenzaamheid | Rijksoverheid.nl. Geraadpleegd op 1 juni 2022, van https://www.rijksoverheid.nl/onderwerpen/eenzaamheid/aanpak-eenzaamheid
17. Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2
18. Pandya, Sharnil & Mehta, Pooja. (2020). A Review On Sentiment Analysis Methodologies, Practices And Applications.
19. Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic Inquiry and Word Count (LIWC): LIWC2007. Austin, TX: LIWC.net.

20. Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Austin, TX: University of Texas at Austin.
21. Pennebaker, J. W., Francis, M. E. F., & Booth, R. J. (2001). Linguistic Inquiry and Word Count. Erlbaum Publishers, Mahwah, NJ. https://www.researchgate.net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC
22. Rapport Taskforce De juiste zorg op de juiste plek. (2018, april). De juiste zorg op de juiste plek.https://www.rijksoverheid.nl/documenten/rapporten/2018/04/06/rapport-de-juiste-zorg-op-de-juiste-plek
23. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135–1144
24. Rosenthal, S., Farra, N., & Nakov, P. (2017, August). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).
25. Sadia, A., Khan, F., & Bashir, F. (2018, februari). An Overview of Lexicon-Based Approach For Sentiment Analysis. d International Electrical Engineering Conference, Karachi, Pakistan. https://ieec.neduet.edu.pk/2018/Papers_2018/15.pdf
26. Salas-Zárate, M. P., Paredes-Valverde, M. A., Rodríguez-García, M. N., Valencia-García, R., & Alor-Hernández, G. (2017). Sentiment Analysis Based on Psychological and Linguistic Features for Spanish Language. Current Trends on Knowledge-Based Systems, 73–92. https://doi.org/10.1007/978-3-319-51905-0_4
27. Spruit, M., Verkleij, S., De Schepper, K., & Scheepers, F. (2022). Exploring Language Markers of Mental Health in Psychiatric Stories. Applied Sciences, 12(4), 2179. https://doi.org/10.3390/app12042179
28. Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. Journal of Personality and Social Psychology, 118(2), 364–387. https://doi.org/10.1037/pspp0000244
29. Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behavior Research Methods, 53(5), 2069–2082. https://doi.org/10.3758/s13428-020-01531-z
30. Van den Broek, A., Videler, A., & Van der Voort, P. (2021, 20 november). Draagvlak creëren voor de ouderenzorg Zes aanbevelingen voor het nieuwe kabine. Gerontologie en Geriatrie. Geraadpleegd op 28 mei 2022, van https://www.researchgate.net/profile/Peter-Voort/publication/357049731_Draagvlak_creeren_voor_de_ouderenzorg_Zes_aanbevelingen_voor_het_nieuwe_kabinet/links/61b9b3ab1d88475981ef0f73/Draagvlak-creeren-voor-de-ouderenzorg-Zes-aanbevelingen-voor-het-nieuwe-kabinet.pdf
31. Vig, Jesse. (2019). BertViz: A Tool for Visualizing Multi-Head Self-Attention in the BERT Model.

8 Appendix

Appendix 1, the form that was sent to the participants to fill in:



Thesis sentimentanalyse onder ouderen

Er worden geen persoonlijke gegevens opgeslagen, alles is dus anoniem.

Bedankt dat je even de tijd wilt nemen om mij te helpen!

Je ziet 3 invoervelden waar je een tekst in kan typen, de bedoeling is dat je hier een tekst neerzet die de week van een ouder persoon beschrijft. Het idee is dat je een overwegend positieve, een overwegend negatieve en een neutrale tekst typt waarbij je je inleeft in een 65-plusser. Zie dit als een soort dagboek gericht op de emotionele status van een persoon gedurende de week. Het gaat hierbij om beschrijvingen over gesteldheid van een ouder persoon (65+). Probeer dit dus ook, voor zover mogelijk, mee te nemen in de teksten. De teksten mogen natuurlijk verzonnen zijn.

Een voorbeeld: "Nou sinds mijn operatie heb ik nog steeds redelijk veel pijn in mijn linker knie, maar ik voel me wel al iets beter dan de afgelopen weken!"

Let op: Het is hierbij belangrijk om zo veel mogelijk in spreektaal te schrijven gezien het eindproject gesproken taal als data zal gebruiken. De lengte van de tekst mag zo lang of kort zijn als je zelf wilt. Varieer hier vooral in!

Als je op "verzend" klikt, zie je de optie om nog een keer deze vragenlijst in te vullen, mocht je hier de gelegenheid voor hebben en mij nog wat willen helpen vul de lijst dan vooral (veel)vaker in!

Ben je geïnteresseerd in het project en wil je hier iets meer over weten? Klik dan op deze link voor een korte toelichting: https://docs.google.com/document/d/1rnNkegjDp3uKC1QgEraRPe3LQLDigAYrSVaYqMsJDT0/edit?usp=sharing

Bedankt!

Met vriendelijke groet,
Rik van Baar

PS: Mocht je vragen hebben mag je me altijd appen of mailen naar: rikvanbaar2016@gmail.com

Overwegend positieve gevoelens *

Voorbeeld: "Euhh vandaag voel ik mij best goed hoor, ik heb minder last van mijn knie en het is lekker weer! Ik ga zo meteen nog even lekker zwemmen en dan komen de kleinkinderen dus dat is fijn"

Tekst lang antwoord

Overwegend negatieve gevoelens *

Voorbeeld: "Het was echt een rotdag en ik ben er helemaal klaar mee!"

Tekst lang antwoord

Gemengde gevoelens(zowel positief als negatief of geen van beide) *

Voorbeeld: "Ik heb afgelopen tijd niks bijzonders mee gemaakt en er is niet veel veranderd. Deze week heb ik ook niet heel erg veel gedaan."

Tekst lang antwoord

Appendix 2: GitHub Link

The GitHub page contains all results, models and other software that is used for this thesis.

https://github.com/Rikvbr/Welzijn.AI_Thesis