EXAMENSARBETE INOM TEKNIK,
GRUNDNIVÅ, 15 HP
STOCKHOLM, SVERIGE 2020

Diagnosis of Dementia using Transformer Models

ALEXANDER ASLAKSEN JONASSON

ALFRED WAHLFORSS

KTH
SKOLAN FÖR ELEKTROTEKNIK OCH DATAVETENSKAP

Diagnosis of Dementia using Transformer Models

Alexander Aslaksen Jonasson
DIVISION OF SPEECH, MUSIC AND HEARING
KTH
Stockholm, Sweden
aljonass@kth.se

Alfred Wahlforss
DIVISION OF SPEECH, MUSIC AND HEARING
KTH
Stockholm, Sweden
alfredwa@kth.se

***Abstract*—Dementia is a syndrome of illnesses resulting in cognitive decline, severely impacting the lives of those afflicted as well as their loved ones. The most common form of dementia is Alzheimer's disease, with roughly 10 million new cases each year. In this study we examine different machine learning models and approaches aimed to aid healthcare professionals in early diagnosis of Alzheimer's disease, potentially automating parts of the diagnostic process. We evaluate our models on the Pitt corpus of the DementiaBank dataset, using 10-fold cross validation. We compare the BERT and RoBERTa transformer models, and find that both models achieve high accuracy, precision, and specificity. The highest accuracy is achieved by RoBERTa, reaching an accuracy of 86.72%, a precision of 90.69% and a specificity of 90.53%. Furthermore, we explore the viability of using automated speech recognition for automatic transcription of audio samples from patient meetings. RoBERTa achieves an accuracy of 83.59% using transcripts generated by Google's automatic speech recognition, suggesting such methods may be viable for automating certain parts of the diagnostic process.**

In addition to the exploration of transformer models and their viability for dementia diagnostics, this paper provides a market analysis of a potential automated diagnostics tool utilizing transformer models. The analysis is based on a literature study and on two interviews; one with the CEO of a start-up providing automated dementia tests for healthcare professionals, and one with a psychologist researching dementia as well as potential methods of early diagnosis of dementia. With the interviews and literature study as a basis, we use the SWOT framework, and PEST analysis along with Porter's five forces framework to analyse the current market potential for such an automated tool. Despite detecting several obstacles and difficulties prior to market entry, we find significant potential for such a product given the current state of the market.

I. SAMMANFATTNING

Demens är ett syndrom av sjukdomar som orskar kognitiv nedsättning och påverkar både de drabbade och deras familjer. Den vanligaste typen av demens är Alzheimers sjukdom, med cirka 10 miljoner nya fall per år. I denna studie undersöker vi olika maskininlärningsmodeller och tillvägagångssätt i syfte att underlätta för sjukvårdspersonal att ställa en tidig diagnos, och möjligtvis att även kunna automatisera vissa delar av diagnosprocessen. Vi utvärderar våra modeller på Pitt-corpuset i DementiaBank-datasetet och använder 10-delad korsvalidering. Vi jämför två transformer-modeller: BERT och RoBERTa, och finner att båda modeller åstadkommer goda resultat avseende noggrannhet, precision, specificitet och sensitivitet. Den högsta noggrannheten uppnås av RoBERTa, på 86.72%, en precision på 90.69%, och en specificitet på 90.53%. Vidare undersöker vi gångbarheten i att använda automatisk taligenkänning för automatiserad transkribering av ljudinspelningar från patientmöten. RoBERTa uppnår då en noggrannhet på 83.59% när den använder transkriberad text från Googles automatiska taligenkänningstjänst, vilket tyder på att sådana metoder kan vara gångbara för att automatisera vissa delar av den diagnostiska processen.

Förutom undersökning av transformermodeller bidrar detta verk även med en marknadsanalys av marknadspotentialen för ett verktyg för automatiserad demensdiagnostik. Analysen baseras på en litteraturstudie och två intervjuer; en med en VD för en start-up som erbjuder liknande tjänster, och en intervju med en forskare inom demens. Med litteraturstudien och de två intervjuerna som grund analyserar vi marknadspotentialen med tre ramverk: Porters fem krafter, PEST-analys och SWOT-analys. Vi fastslår att det trots flertal hinder och svårigheter för marknadsinträde finns det stor potential och en stor efterfrågan på en sådan produkt.

Part I
Transformer models for early diagnosis of dementia

I. Introduction

A. Background

1) Dementia: Dementia is a syndrome which results in cognitive decline, affecting areas such as thinking, memory, language, judgement and orientation. It is the result of several diseases, including Alzheimer's disease (AD), dementia with Lewy bodies and frontotemporal dementia. AD is believed to constitute roughly 60-70% of all cases worldwide. A total of 50 million people are believed to suffer from dementia worldwide, with roughly 10 million new cases each year, 10 000 - 15 000 of these being in Sweden. Somewhere between 5-8% of the worldwide population aged above 60 suffer from dementia. The economic implication of dementia is believed to be around $1 trillion, or roughly 1.1% of the world's gross domestic product. The prevalence is expected to increase significantly in the near-future considering the world's rapidly aging population and low birth-rates [30].

Unfortunately, there are few treatments available against dementia. However, the potential medication and remedies that do exist have been shown to be most effective if implemented early (Posner et al., 2017). As a consequence, it is of high

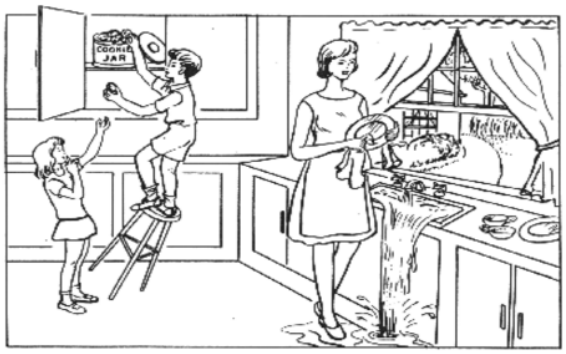

Fig. 1. Cookie Theft Picture, adopted from Goodglass et al. [6]

importance that there are cost-effective screening methods that can detect and diagnose dementia from early-signs. The current methods used for diagnose dementia early uses either positron emission tomography (PET) or magnetic resonance imaging (MRI). Both are expensive but non-invasive (Nensa et al., 2014).

2) Dementia's effect on speech: There is considerable evidence that dementia, specifically of the Alzheimer's type, affects speech. Alzheimer's disease patients score significantly lower than the controls in the areas of verbal expression, auditory comprehension, repetition, reading, and writing [10]. Szatloczki et al. connects speech tempo, pauses in speech, and speech length to early stages of the disease [25]. This shows that speech could be analyzed in order to diagnose the disease early.

Hesitations, silences and filler words such as "ehm" and "uh" are more likely to occur during conversations for persons with dementia, as they more frequently forget details and context of the conversation. Khodabakhsh et al evaluated both linguistic and prosodic features for Alzheimers detection in speech [15]. They conclude that prosodic features are superior to linguistic features when it comes to detection [15]. Khodabakhsh et al found features such as silence to utterance ratio, response time, average word count, word rate, and filler word rate to be useful features for classification [15].

Analysing the speech output of patients could therefore be a potential tool for creating effective diagnostics tools. One such test that uses the speech of patients in order to screen for Alzheimer's disease is the Boston Cookie Theft Picture Description Task 1. In the Cookie Theft test, the patient is asked to describe everything occurring in the picture. The picture itself includes details and information of various levels, containing different semantic categories as well as causal and temporal relations between objects. The complexity of the depicted situation allows for thorough verbal and speech analysis of the patient describing it.

This test currently requires the expertise of specialised medical doctors. However, with the advent of natural language of processing there is the possibility to automate this test. Such an automation would speed up current waiting times, and lessen the burden of the health care system in general. Furthermore, democratising an early diagnose could potentially extend the life-span of millions of dementia patients. Numerous research groups have worked on automating this test, with varied results [19] [3].

3) Transformers and Natural Language Understanding: The publication of "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" brought forth a new paradigm for natural language processing [9]. It presented a new architecture for natural language processing that uses transformers and it has surpassed the state of the art results in all major text classification tasks. The main benefit is that the model can be pre-trained on large amounts of data and then fine-tuned to fit the specific task. This is possible because the new models are bidirectional, i.e. they can incorporate context from both directions at the same time.

In this paper, we apply the new transformer architecture to the Cookie Theft picture task. We want to investigate whether the transformer architecture can improve the current results and achieve a state of the art accuracy for diagnosing Alzheimer's disease using language form the Cookie Theft picture task.

In RoBERTA: A Robustly Optimized BERT Pretraining Approach researchers at Facebook AI and University of Washington show that BERT was severely undertrained [17]. They propose a new training method for a new model referred to as RoBERTa. They modified the BERT training process in several ways, including increased training time, larger batches and data sets, changing masking pattern during training dynamically and using longer sequences for training. They also skipped the objective of next sentence prediction during training. With these modifications they were able to achieve SoTA results on a variety of tasks, outperforming the original BERT on several of them.

II. Previous Work

The papers that use machine learning in order to diagnose AD with dementia bank differ in a few ways. First, some use the transcripts from the dataset while others apply automatic speech recognition to transcribe the speech audio into text. Second, the studies use different language models. Thirdly, some include the data from those with mild cognitive impairments and others do not. Finally, they have a variety of evaluation metrics.

While there are many papers published using Dementia Bank, we will discuss four papers using different methods for making an automatic dementia diagnose with Dementia Bank: Zhou et al. used a SVM classifier from text features [31], Wankerl et al. utilizes a n-gram model [28], Hernández-Domínguez et al. applied a SVM and random forrest classifer with phonetic and linguistic features [13], Guo et al. created an algortihm that uses the perplexity feature of a n-gram model [11]. Now we will give an overview of the main differences between them.

A key difference is whether to use the provided transcripts or to use automatic speech recognition in order to transcribe

speech into audio. The results seem to be better when using the transcripts, which makes sense intuitively since the transcripts will be more accurate than anything generated with automatic speech recognition. However, using transcripts outside of the lab is prohibitively expensive. Therefore, it would be more interesting to use automatic speech recognition in our study.

Another important aspect is which models the papers use. Two studies use n-gram models [28] [11]. Wankerl et al. uses two trigram models using the transcriptions and they derive a single feature by calculating the difference between the perplexities in the two models [28]. Guo et al. uses a two-dimensional perplexity feature which is combined with some baseline features [11]. Hernández-Domínguez et al. has an SVM and a Random Forest Classifier, where the Random Forest Classifier performs the best [13]. Zhou et al. also uses an SVM [31]. No study has so far applied the new transformer models, such as BERT, to the DementiaBank corpus.

It is difficult to precisely define the state of the art since each paper uses its own evaluation metric. Zhou et al. uses 10-fold-cross-validation but never explicitly states the diagnostic accuracy, instead, they focus on the word error rate of their automatic speech recognition [31]. Hernández-Domínguez et al also uses 10-fold-cross-validation with a division of 10% test data [13]. They report an average accuracy of 87%; however, it is unclear whether they divide the data between individuals. Thus, they might be training on data samples of the same individuals which means they overfit to the data. Wankerl et al. evaluate their results with leave-one-out-cross-validation, which means that the data is divided into one part for every person [28]. The model is trained on all data, except for the data from one person and then the model is tested on that person. They use an equal-error-rate which gives them an accuracy of 77.1%. Furthermore, since some studies choose to not use the data from subjects with mild cognitive impairments while others keep them, it is difficult to determine the actual state of the art accuracy. A final point which further complicates the comparison between papers is whether they use the provided transcripts or not. Using the transcripts most likely increases the accuracy.

III. Theory

A. Dementia and speech

1) Cookie Theft Picture: The Cookie Theft test is a common part of dementia evaluations and consists of the patient describing the Cookie Theft Picture 1. The picture was first included in the Boston Diagnostic Aphasia Examination [6], and level of detail in a person's description of the picture varies greatly depending on the person's cognitive abilities.

The Cookie Theft picture is a widely used task in clinical settings for evaluating cognitive and verbal abilities of patients suspected of having cognitive disorders. The patient is asked to describe everything occurring in the picture. The picture itself includes details and information of various levels, containing different semantic categories as well as causal and temporal relations between objects. The complexity of the depicted situation allows for thorough verbal and speech analysis of the patient describing it.

B. Transformers and Natural Language Understanding

With the publication of "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", a new paradigm for natural language processing was unleashed [9]. They presented a new architecture for natural language processing that uses transformers and it has surpassed the state of the art results in all major text classification tasks. The main benefit is that the model can be pre-trained on large amounts of data and then fine-tuned to fit the specific task. This is possible because the new models are bidirectional, i.e. they can incorporate context from both directions at the same time.

C. Transformer architecture and Attention Mechanisms

For a long time recurrent neural networks (RNNs), with various architecture including long short-term memory (LSTM) and gated recurrent units (GRU) were the dominating architecture within natural language processing, sequence and language modeling. These were able to achieve state-of-the-art results on various tasks, but performance is oftentimes lowered for longer sequences. Bengio et al have previously demonstrated the difficulty of RNNs to capture long-term dependencies [14]. Consider a sentence such as "The blue ball rolled across the field as it slowly made its way towards the goal". In this sentence it is easy relatively simple for an RNN to capture that "it" refers to the ball, but for longer sequences containing several sentences, such dependencies are tricky to capture. Several attempts were made to improve the performance, such as was the LSTMs and GRUs, which both are RNNs modified to capture longer-term dependencies.

Another attempt to has been through the mechanism of Attention, which enabled models to take more context and dependencies into account, regardless of their distances within the sequence. Attention mechanisms were typically incorporated into RNNs. RNNs are sequential in nature, and require sequential computations, which can be time consuming and do not allow for higher degrees of parallelization.

In above sentence with the ball, the self-attention mechanism allows the transformer to model the relationship between every token in the sequence, and thus deducing that the word "ball" refers to spherical ball, as in football, rather than a masquerade ball. [1]

In their seminal paper "Attention Is All You Need", researchers at Google introduced a novel architecture for sequential modeling, free from recurrence, relying instead solely on attention mechanisms, which allow for more parallelization than RNNs and previous architectures. Below follows an overview and description of the architecture. [27]

D. Architectural Overview

Transformers can be used for a wide variety of tasks, including next-sentence prediction, named entity recognition, classification tasks such as sentiment analysis, and neural language translation. This section describes the transformer

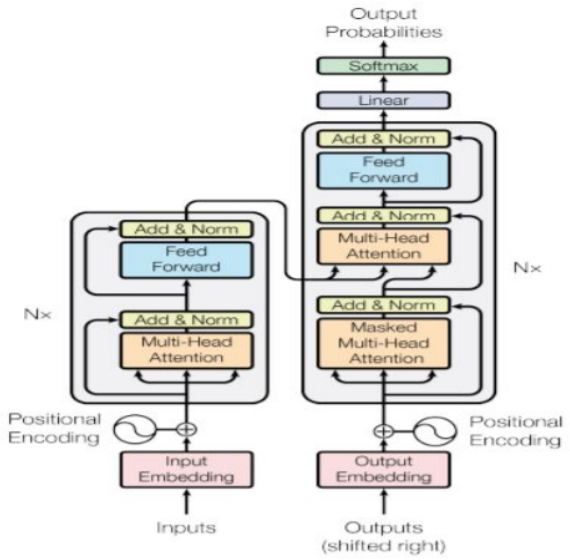

Fig. 2. Transformer Architectural Overview, adopted from Vaswani et al. [27]

architecture for language translation, as it is presented in the seminal paper Attention is All You Need. The architectures for other tasks are similar, but the final layer after the decoder stacks may differ. Fig. 2 displays the overall architecture.

The input sequence is transformed into a sequence of embedding vectors. These are then fed to the first out of 6 encoders. The first encoder's output is fed into the second encoder and so on and so forth, until it reaches the 6th and final encoder. Each enconder consists of a self-attention layer and a feed forward layer.

There are 6 decoders in total, each consisting of a self-attention layer, followed by an encoder-decoder attention layer, and finally a feed forward layer. Following the last layer there is a fully-connected layer and a SoftMax layer, finally yielding an output token. The input sequence is fed only once to the encoder stack, and the outputs of the encoder stack are then fed to the decoder stack every time step. The output of the decoder stack is also fed to the bottom decoder for each time step. This process is performed until the transformer finally outputs an end-of-sequence token. [27]

We will now more thoroughly examine each component of the architecture.

1) Word Embeddings: Each word is transformed into a word embedding vector, representing semantic features.

2) Encoder and Decoder: Like many other successful architectures, the Transformer uses an encoder stack, and a decoder stack, each consisting of 6 identical layers.

3) Attention and Self-Attention: Attention can be described as a function which maps a query and a set of key-value pairs to an output. [27]

From the vector embedding of each input token in the sequence, a Query vector, Key vector, and Value vector is created. These are generated by multiplying the word embedding vector by one matrix (whose weights are learned during training) each. These have a dimensionality of 64, while the embedding vectors have a dimensionality of 512. The self-attention itself is a vector.

Assume that the input consists of a sequence of 3 tokens. Let x_i, q_i, k_i, v_i represent the word embedding, query, key, and value vectors of token i in the sequence. To calculate the self-attention vector z_1 of the first token, the following calculations are performed. First, the dot-product between q_1 and k_i is calculated for every token i in the sequence. This score is then divided by the square root of the dimensionality of the q, k, v vectors, in this case 8. This is done in order avoid the values becoming too large, which could result in low gradients due to pushing the score far to the edges of the softmax function. After dividing the score, the softmax function is applied to each score. This new score is then multiplied by the v_i for each token in the sequence. Finally these value vectors multiplied by softmax are added together, resulting in z_1.

In summary:

$$z_i = \Sigma_{j=0}^{N} Softmax(\frac{q_i \cdot k_j}{\sqrt{d_k}}) * v_j$$

where N is the length of the sequence, and d_k is the dimensionality of the q, k, v vectors.

This process is made computationally efficient by using matrix multiplication as follows [27] [24]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V.$$

4) Multi-headed attention: The model contains several sets of Query, Key and Value matrices for every encoder and decoder, allowing for several representation subspaces. The multi-headed attention mechanisms increases the model's capacity to attend different representational subspaces at several positions of the sequence, in parallel.

The encoder/decoder layer however requires a single matrix, not several from each attention head. Thus, the matrices $Z_i, ..., Z_n$ are concatenated so that $Z_i, ..., Z_n = Z$, and then multiplied by a matrix W^O whose weights are also learned during training.

In matrix format, it can be formulated as follows:

$$MultiHeadAtn(Q, K, V) = Concat(head_1, ..., head_n)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where n is the number of heads, set to 8 in the original paper [27] [24].

5) Positional Encoding: As mentioned previously, the transformer does not use recurrence nor convolution, so in order to account for the order of the tokens in the input, information regarding absolute as well as relative positions of the tokens is added.

This information is stored in a positional encoding vector, of the same dimensionality as the word embedding vectors, i.e. 512, in order to be able to add them together. The idea is that meaningful information regarding positions is provided

by adding them to the embeddings, by altering the distances between the word embedding vectors when they are used to obtain the attention, as well as when they are projected onto the query, key, and value vectors.

The transformer uses a sine and cosines as positional encodings. The formula is given by

$$PostionalEncoding_{pos,2i} = sin(\frac{pos}{10000^{2i/d_{model}}})$$
$$PostionalEncoding_{pos,2i+1} = cos(\frac{pos}{10000^{2i/d_{model}}})$$

where i is the dimension, while pos is the position. [24] [27]

6) Residuals: Each sub-layer in the encoders and decoders contains residual connections as well as normalization steps.

For encoding layer 1, the procedure would be as follows:

$$x_1 + positionalencoding(x_1) = x_1'$$

This is then fed to the self-attention layer: $SelfAttention(x_1') = z_1$
These are then added, normalized, and fed to the feed forward layer:

$$FF(normlayer(z_1 + x_1'))$$

The output of the first layer-norm and the feed forward layer are then added together and normalized before being passed on to the next encoder [24] [27].

7) Layer Normalization: Layer normalization helps speed up training time and is similar to batch normalization, with a few key differences. While batch normalization calculates mean and variance in order to normalize the input values to a neuron over a batch of training samples, reducing the training time of the neural network, this is not applicable to RNNs. Layer normalization calculates the mean and variance not over a batch of training sample, but of the summed inputs to a single layer during on one training sample. Also, layer normalization performs the same computation during training as well as during testing. Furthermore, it can be applied effectively to RNNs [4].

8) Decoder: The decoders are similar to the encoders, but there are a few differences. The encoder is used only once, while the decoder is used for several time steps until an end-of-sequence token is outputted. For each time step, the output of the encoder is used as input to the encoder-decoder attention layer of every decoder in the stack.

Each decoder contains a self-attention layer, an encoder-decoder layer, and finally a feed forward layer (and also residual connections and norm-layers).

The final encoder in the sequence is transformed into two vectors K, and V, which are fed into the encoder-decoder attention layer of each and every decoder layer. This allows the decoder to pay attention to the relevant positions in the input sequence. The final decoder layer passes its output into a linear fully-connected layer, which is then softmaxed and produces an output (described further in a later section).

The decoding process continues for several time steps. The output of every time is embedded, and positional encoding is added, and the product is concatenated to previous outputs (positions not yet seen are masked with $-\infty$), and the product is fed into the first decoder. This process continues until the final layer outputs an end token, signifying the end of the sequence [24] [27].

9) Final layer, Softmax: The final decoder layer yields a vector containing floating-point values, which are fed into a fully-connected layer which produces a large logits vector. The logits vector has the same dimensionality as the number of tokens in the vocabulary. If there are 100 000 unique tokens in the corpus, then the logits vector has 100 000 entries. The softmax function is then applied to the logits vector, resulting in a vector of probabilities, where each entry is positive and adding each entry sums to 1.0. The position of the entry with the highest probability is noted, and the token associated with this entry is then the token outputted by the transformer [24] [27].

10) Beam Search: The transformer can be considered to output a probability distribution over all the tokens in the vocabulary, where the token corresponding to the entry with highest probability is the token the transformer considers most likely. One method is to always choose the entry with highest probability as the output token, but there are other methods. One popular method is known as Beam Search. After the first step, the model remembers the top n words, and runs the model n times during the second step, one time for each of the n tokens, letting the model act as if there are n potential sentences. This process is repeated for the following steps, and the sequence that overall accumulated the largest probability is chosen. Several parameters can be manipulated. Beam size refers to the number of steps in which token candidates are considered, in the above example the beam size is 2 because results were compared after computing beams for the first and second time steps. Top beams refers to the number of tokens considered, which in the example was n [24] [27].

IV. Method

A. Data

The dataset used for prediction consists of the audio recordings of the Cookie Theft test from DementiaBank's Pitt corpus and their corresponding transcripts. As mentioned above, the Cookie Theft is a natural choice for dementia detection of continuous speech samples. The samples contained in the Pitt corpus are taken from a large longitudinal cohort study of AD conducted between 1983 and 1988. After removing participants who either developed dementia during the study, or showed to have other diseases affecting cognitive abilities, 188 (out of 204) participants had definite or probable AD, and 101 (out of 102) participants were in the control group. The participants had to fulfill certain criteria in order to be eligible for the study, including not having had any previous cognitive disorders, nor having had any medication affecting the central nervous systems (excluding antidepressants). The participants underwent several sessions of medical and cognitive testing. Thus, a total of 289 participants were included in the study. The corpus contains a total of 306 (as some

participants performed the test more than once) audio samples and their corresponding transcriptions, transcribed by linguists, complete with parts-of-speech tags [5].

B. Data Processing

Whereas transformers have shown to be extremely powerful tools when it comes to understanding and processing language coming from written text, they are not designed to process audio. While many clues concerning the speaker's cognitive abilities are likely to be found within the transcription, they do not provide the entire picture. As mentioned above, dementia affects many aspects of speech, including enunciation, speech production rate and other possible prosodic aspects. To make use of these potentially insightful features, we use speech analysis software provided by PRAAT to extract syllable intervals as well as fundamental frequency (f_0) sequences from the audio files. Furthermore we use the OpenSMILE library to obtain over 5000 additional features from each sample.

1) Automatic Speech Recognition: As mentioned above, the dataset contains both audio files recorded during the participant meeting as well as written transcripts from these meetings transcribed by linguists. BERT and similar NLP models typically require text input, and in order to make a diagnostic aid to help healthcare professionals make a diagnosis, it is highly beneficial if as little manual labor as possible is required. Rather than writing transcripts by hand, automatic speech recognition (ASR) could be used to transcribe the speech from patients during the meeting. The performance of ASR models has increased greatly in recent years, and in this study we attempt to use three variants of Google's ASR service, one in which the audio is inputted as is, which we refer to as *Low*, one in which the volume has been increased significantly, referred to as *High*, and one in which the audio is inputted as is, but the model used is an enhanced model provided by Google, referred to as *Enhanced*.

C. Bert Embeddings

Simply using BERT for classification will yield a single value; 1 or 0 depending on the classification. To combine the output of BERT with other non-text features, we can use BERT to get feature vectors. Devlin et al. [9] describes different variations of combining the last layers of BERT as contextual embeddings, and then feeding these as input to a BiLSTM before the classification layer in a Named Entity Recognition (NER) task. Using a concatenation of the last four hidden layers (out of 12) achieved the best results for the NER task. This method achieves an accuracy of 96.1%, and a weighted sum of the last four hidden layers achieves an accuracy of 95.9%, while using the entire BERT-base model achieves an accuracy of 96.4%. This shows that this approach can achieve comparable results to using the entire fine-tuned BERT model [9]. For our study, we use the method of concatenating the last four hidden layers, and then concatenating them with other prosodic and linguistic features, before applying a final classification layer.

D. General Description

Transcriptions are used as input to a transformer model in order to extract embedding vectors by concatenating the four final hidden layers of the output. The PRAAT software is used on the audio samples to produce f_0 sequences and syllable intervals, which are themselves used as input to an LSTM model. An additional number of features are also extracted from the audio samples using OpenSMILE. These features are inserted into a vector. The BERT embeddings vector, the LSTM vector, and the openSMILE features vectors are finally concatenated and used as input to one or more fully-connected layers, yielding a binary output as classification.

BERT has several settings, and maximum sequence length the model can take as input can be adjusted. In our experiments, we tried to set the maximum sequence length to both 256 and 512 tokens. If a sequence contains more than the limit, only the first tokens up until the limit are included, the rest are discarded. If the sequence contains fewer tokens than the limit, the input vector is padded until it is of the appropriate length.

1) Data exploration and general comments: 10-fold cross validation was used to reduce bias. Each patient in the study participated in the cookie theft test between one and three times, thus, in order to avoid training a model on one sample from patient A, and also testing the model on another sample from patient A, the samples were grouped together on a per patient basis. As such, the data was divided into train-test splits according to patients rather than samples. After removing patients classified as MCI, 275 patients were left with a total of 512 data samples. The 275 remaining patients were split into 10 sets, averaging 27 patients per set. For each train-test set a new model was trained for 8 epochs, and the highest number of correctly classified samples was recorded. The total number of correctly classified samples spanning over all train-test sets was finally recorded and accuracy, precision, recall and specificity was calculated.

For all models with maximum length set to 256 a batch size of 16 was used, while for models with maximum length set to 512 a batch size of 6 was used due to GPU constraints. We did not experiment thoroughly with different hyperparameters, instead we used many of the default settings. A learning rate of 3e-5 was used, with the AdamW optimizer from the HuggingFace library.

We noted that the all models seemed to struggle with roughly the same data samples. The visits of three patients were not correctly classified by any model.

Out of the 10 most often incorrectly classified patients, 3 belonged to the control group and 7 belonged to the AD group. In total, there are 269 AD samples and 243 control samples, meaning that 52.5% of the samples are AD. There is a total of 275 patients, 176 in AD group, and 99 in control.

E. Metrics

We employ four different metrics; accuracy, precision, specificity and recall. They are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN},\ Precision = \frac{TP}{TP + FP}$$
$$Specificity = \frac{TN}{TN + FP},\ Recall = \frac{TP}{TP + FN}$$

V. Results

A. Results Overview

1) Prosodic Features with Bert Embeddings: We found that concatenating the prosodic features with the embeddings of BERT and RoBERTa and then feeding them to a fully-connected layer led to a slight decrease in accuracy for all models, and as such these are not included in the table of results and were not analyzed further.

TABLE I
BEST RESULTS IN TERMS OF ACCURACY FOR EACH MODEL ON LINGUIST TRANSCRIPTS. 256 AND 512 REFERS TO THE MAXIMUM INPUT LENGTH USED.

Model	BERT256	BERT512	ROBERTA512	Guo et al. [11][1]
Accuracy	84.96%	85.55%	**86.72%**	85.4%
Precision	85.82%	84.45%	**90.69%**	
Specificity	84.36%	81.89%	**90.53%**	
Recall	85.50%	**88.85%**	83.27%	

TABLE II
BEST RESULTS ON DIFFERENT TRANSCRIPT TYPES.

Transcript	Linguist	High	Low	Enhanced
Model	RoBERTa512	BERT512	BERT512	RoBERTa512
Accuracy	**86.72%**	82.23%	81.84%	83.59%
Precision	**90.69%**	82.48%	83.85%	86.56%
Specificity	**90.53%**	80.25%	82.71%	86.01%
Recall	83.27%	**84.01%**	81.04%	81.41%

prediction outcome

actual value	**p**	**n**	**total**
p′	True Positive: 224	False Negative: 45	P′ = 269
n′	False Positive 23	True Negative 220	N′ = 243
total	P = 247	N = 265	

TABLE III
CONFUSION MATRIX FOR HIGHEST PERFORMING ROBERTA512 MODEL

[1]Guo et al. uses leave-one-person-out validation instead of 10-fold-cross-validation

TABLE IV
SAMPLE OF THE SENTENCES MOST OFTEN INCORRECTLY CLASSIFIED.

Most often incorrectly classified sample in AD group	well the boy on the chair is falling, reaching up for a cookie, handing one to the girl. the lady is wiping a dish. water running on the floor. she's standing in it. trees outside, the lawn, shrubbery. a window outside that I can see. that's about it dear.
Second most often incorrectly classified sample in AD group	the chair is tilting. lid is off of the cookie jar. cookie in the left arm of the boy. his right hand is touching a cookie in the cookie jar. one of his feet is a about a third off of the stool. he's got short pants and a blouse. and look like they're boots instead of shoes. the girl has a finger to her lips as though to say "quiet" one hand out. her left hand is out. she's got short skirt and a blouse, jersey sweater. socks, anklets rather. the on I did mention that the stool was tilting. the boy's standing on the stool and it's tilting. I think I mentioned that. the girl has hair hanging to her neck. the boy has like hair combed straight back. the jar is open on the the the cupboard. the mother's drying the dish with her right hand holding it with her left. she's got an apron over her dress or whatever it may be. water spilling out of the sink. two cups facing opposite direction. one plate to the right of the cups. curtains flowing in the breeze of the wind. there still some more but oh yeah.
Most often incorrectly classified sample in control group	inside the room or every place?. *INV just say it out loud. oh you don't want me to memorize it ! oh. okay, the the little girl asking for the cookie from the boy who's about to fall on his head. and she's going I guess "shush" or "give me one". the mother's we don't think she might be on drugs because she's off someplace because the sink's running over. and it's summer outside because the window's open and the bushes look healthy and she's drying dishes with her apron on. and the cookie jar's looking full. that's it.
Always diagnosed correctly and belonging to AD group	there's a little girl talking to this boy up up on the step. and she's asking him to bring some of this down or whatever it was a jar or whatever it is that so it doesn't doesn't break. there's a mother over here. she's watching them that she has that break in her hand. and it looks like it's very strong or heku or she's looking real real good at it, a jar or whatever. she's touching giving a little little little touch of her mouth. although that's. and it it looks like he's gonna bring some of that down down for them all. and and and mother's out there looking at them. and and looks like she's washing this dishes that they already had it. and she's she's washing the dishes away from them. it it looks that way. she's cleaning it you know.
Always diagnosed correctly and belonging to Control group	cookie jar. a lad standing on a stool teetering, grabbing for the cookies. sister I guess laughing at him. mother washing dishes. sink is overflowing. view of the yard and the kitchen window with its curtain. two cups and a dish remain. looks like they're dried. mother standing in the overflowed water. her two faced cabinets four doors. and a valance and the curtain. window's half open. and there's landscaping along the wall of view of the yard. and the walk pictured from the window. water is running, overflowing. boy is holding a cookie in his left hand, grabbing for another one with his right hand. the lid of the cookie jar is over uff off. mother's wearing an apron drying dishes. okay.

TABLE V
SAMPLE TOKEN AND CHARACTER LENGTHS AND STATISTICS

Group	All	AD	Control	Worst 10
mean no. chars.	502.89	466.77	543.01	520.73
median no. chars.	446	420	475	481
std.dev. no. chars.	265.93	251.68	275.41	206.13
mean no. chars.	128.26	121.87	135.35	131.55
median no. tokens	114	109.5	120	126
std.dev. no. tokens	65.39	64.08	66.10	49.51

VI. ANALYSIS

We find that the transformer architecture is suitable for classification of text samples taken from cookie theft tests. All models consistently significantly perform well above the baseline, which we define as predicting the most common category on all samples. The RoBERTa model achieved the highest accuracy, precision, and specificity, which were all achieved using the linguist transcripts. The ratio between correctly classified AD samples (224) and correctly classified control samples (220) is roughly 1 for the best performing RoBERTa model, not displaying any significant bias towards either group. However, we note that the model is more likely to incorrectly produce false negatives (45), compared to false positives (23). Table III shows the differences between various statistics regarding number of characters and number of tokens in the different groups. We note that on average the AD group produced samples with fewer total number of characters and tokens (where "tokens" refers to tokens as produced by the BERT tokenizer included in the BERT model). We note that for the group of the 10 least correctly classified patients (referred to as Worst 10 in the diagram), the mean number of characters and tokens was slightly above overall mean.

BERT models overall generally performed slightly worse than RoBERTa, but was able to achieve a higher recall in some instances, with BERT512 achieving the highest recall of 88.85% on the linguist transcripts.

An analysis was made comparing which participants and their samples each model was able to correctly classify. We found that all models were able to correctly classify a large number all samples of a certain group of participants. These samples were correctly classified across all models (BERT256, BERT512, RoBERTa512). Furthermore, we found that there were a few patients whose samples no models were able to classify correctly. Thus we conclude that all tested models struggle with roughly the same set of samples.

Observing the sentence samples in table IV, we note that the samples always correctly classified are rather clear as to whether they belong to the AD group or the control group. Meanwhile, observing the most often incorrectly sample from the Control group, we note that it includes a bit of gibberish, and the authors (admittedly lacking medical degrees) would spontaneously classify this sample as belonging to AD. While this analysis is done ad-hoc, it lends extra credibility to the results of the transformer models.

Unsurprisingly we find that all models performed better when using linguist transcripts, and that the performance goes in the following order: Linguist transcripts, google enhanced transcripts, google high volume transcripts, google low volume transcripts. However, we note that the performance is still relatively good, and not significantly lower than using the transcripts. The lower performance, however, is likely due to errors when transforming the raw speech data to text by Google's ASR model. Another source of error is the fact that the Google transcripts contain words and phrases uttered by both the doctor and the participant, while the linguist transcripts only contain phrases uttered by the participant. Modern microphones commonly have the ability to detect speaker direction, making it easy to automatically remove sound coming from one source or speaker, removing this source of error. The fact that the audio of the doctor is included could be viewed as a feature, rather than a source of noise, depending on the application. A medical examiner may through experience notice subtle details and probe the patient in certain ways to gain useful information. Including these probes could increase the accuracy of a model, and could be useful if including the transformer model as a biomarker in a system of tests. It is also important to bear in mind that the recordings are relatively old and of rather low quality, Google's ASR is likely to produce better results with samples recorded with better equipment.

Taking these things into consideration, an automated setup in which modern equipment is used, removing the instructor's voice from the sample, and then using a model similar to Google's ASR to produce transcripts is likely to yield results of high accuracy similar to those we were able to produce with RoBERTa using the linguist transcripts.

1) Sources of error and improvements: We note that the dataset is relatively small and may not be large enough to fine-tune a transformer model to its full potential. Furthermore, the limited dataset size may affect the reliability of the accuracy, as the train and test split may affect the outcome significantly. However, using 10-fold cross validation strengthens then reliability of the results. Several studies use leave-one-out testing, however, due to limited resources and the time required to train and test each model, we were unable to perform such an analysis.

In recent years a multitude of successful NLP models have been published, including XLNet, TransformerXL and GPT-2, which all have achieved impressive results on several NLP tasks. With more time and resources, we would have like to explore the performance of several of these on the dataset. As mentioned in the results, using prosodic features did not yield any improved results. We speculate that this may be due to the low quality of the audio recordings, and it would be interesting to see if higher quality audio could yield any improvements. Finally, we could have further experimented with hyperparameter tuning, performing a more extensive analysis of how varying the learning rate, batch size, different learning rate schedules and such could affect the performance.

Part II
A market analysis of early dementia screening using natural language processing

I. Introduction

This part of the paper focuses on a market analysis of early dementia screening tools. We will analyze the market potential for a dementia diagnostics system that uses natural language processing. Specifically, we will discuss questions such as: what is the market size and growth for early dementia screening, who are customers, and more. This analysis will be based on Porter's five forces, PEST and the SWOT framework. The section consists of these parts: methodology, explanation of the theoretical framework and, finally, application of the framework on a product that diagnoses dementia early.

II. Methodology

Firstly, an extensive literature study was performed which focused on finding the benefits and problems with early dementia diagnostics. Furthermore, we researched the cost of dementia to society and to the individuals who are affected. This literature study was performed by searching for peer-reviewed articles using the key words: dementia, market analysis, early diagnosis, cost of illness, dementia, Alzheimer's disease, meta-study. Using this search strategy, we screened 153 peer-reviewed articles based on abstract and titles. This screening was mainly focused on finding qualitative meta-studies, since those studies have themselves screened thousands of papers.

Secondly, we had a semi-structured interview with two experts in the field of dementia diagnostics: the CEO of Mindmore, developer of the leading dementia screening tool in Sweden, and a healthcare professional researching potential automated methods of early diagnosis of dementia. These two interviews gave a holistic perspective on the market for dementia diagnostics and was a good complement to the literature study.

The facts discovered in the literature review and the interviews was then arranged in Porter's five forces, PEST and SWOT framework in order to organize the key findings.

III. Theoretical framework

We use a multi-step analysis consisting of the combination of the most used and well respected frameworks for market analysis: Porter's five forces, PEST-analysis, and SWOT. Porter's five forces is a framework used to analyse the the most important market forces [21]. PEST is a wide framework that encompasses political, economic, social and technological factors [7]. SWOT is a general framework that makes it easy to structure information [12]. In each framework we use data from the literature study and the expert semi-structured interview. This novel combination of these three frameworks enables us to use to get the best utility from all three frameworks. The combination is greater than the sum of its parts since the frameworks complement each other in an efficient manor. First, we use Porter's five forces to breakdown the general market conditions. Second, the PEST analysis is applied in order to widen the analysis to include socio-political factors. Finally, we use the SWOT to pick up remaining parts of the analysis which cannot be accommodated by Porter's five forces and PEST.

IV. Result

V. Porter's Five Forces

This analysis is based on the answers provided by the CEO of Mindmore.

A. Threat of new entrants

Entry barriers are relatively high due to most customers requiring several scientific studies displaying the reliability, specificity, and cost efficiency of the product before considering trying it. Capital requirements are however relatively low as no expensive equipment is required to build or develop the product, but funding research to prove its viability may be challenging. Due to these factors, the threat of new entrants is low.

B. Threat of substitutes

The product itself is a major substitute to existing, traditional methods of dementia diagnostics. The method itself could potentially provide significant efficiency in terms of time and cost, as well as reliability and specificity when compared to the traditional methods. Furthermore, the method relies on cutting-edge technology and the reliability of computers, rather than traditional pen-and-paper methods in which results may vary greatly depending on the person in charge of performing the test. Switching costs are relatively low, as the traditional methods for which our product acts as a substitute generally do not require expensive hardware, and new technologies require relatively inexpensive tools. The propensity for a customer to buy could also be increased, as many government goals involve increasing electronic healthcare services and utilizing new technology. Therefore the interviewee from Mindmore deems the threat of newer substitutes low, but there exists a threat in the sense that several actors may want to continue using traditional methods rather than adopting new technology.

C. Bargaining power of customers

The customers in this case refers to several potential groups. Dementia diagnostics is carried out at university hospitals, regular hospitals, outpatient clinics (both privately owned and public), psychiatrists (where access to regular hospitals and outpatients clinics is scarce), and specialized dementia and memory clinics. The interviewee further states that the patient should be viewed not as a customer, but as a partner. The method of selling the product differs within different countries, and even within different regions. Some areas may require the purchases to be done through public procurements, which are typically extensive, whereas private actors may proceed

with greater speed. The bargaining power of customers is high in the sense that significant research and contemplation is performed before buying, given that the product affects the health of patients. Customers may require significant scientific evidence for the specificity, reliability and general utility of the method before considering abandoning the prior methods of diagnosis. However, the interviewee from Mindmore states that studies have been made displaying both reliability and specificity, and several trial runs have indicated a great increase in efficiency regarding speed, reliability and quality control regarding diagnosis, oftentimes greatly reducing the burden many outpatient clinics are facing, given that between 1-2% of all patient visits to outpatient clinics are due to dementia.

D. Bargaining power of suppliers

While Mindmore uses software from a multitude of suppliers, there are relatively many different suppliers for similar tools, and there is also ample opportunity to create this software in-house if suppliers should increase prices or discontinue their software. As such, the bargaining power of suppliers is considered low, and not a potential obstacle to market entry.

E. Competitive rivalry

There are currently several other actors (although most of them in early stages) on the global market, ranging in size from small to large and listed on stock indices. These include Cambridge Brain Sciences, a company offering online cognitive assessments. While competition may be difficult with larger established actors, these have low market penetration in large parts of the world and are mostly locally active. In Sweden, there are currently two major actors; Mindmore and Geras solution. However, currently these companies are quite differentiated and their offerings are far from identical, and competition cannot be considered particularly fierce, especially considering the vastness of the market, and the wide range of interested parties. In summary, rivalry may become a larger problem in the future, but is at this stage considered relatively low threat.

VI. PEST

A. Political factors

Since the likely customer for a screening tool for dementia is the national and regional governments, the political factors are of high importance. Both of our interview subjects mentioned the importance of ”Socialstyrelsen” in Sweden, which is the national agency that sets the guidelines for all health care providers in Sweden. This agency is ultimately controlled by politicians which makes the market more volatile, since they might shift opinions if a new party wins an election. In the interview with the health care professional, there are two things which is needed in order to convince politicians: scientific proof of economic savings and proof that patient's lives are greatly improved.

An issue that might complicate the political landscape according to the CEO of Mindmore is that politicians might be affected by the acts of other digital health care tools. In Sweden, there has been some critique of the private telehealth providers and that critique might have some spillover on other digital tools in health care. Politicians are in general unreliable and might quickly change their mind if some big scandal occurs.

B. Economic factors

Dementia patients take a large economic toll on the health care system. Those with severe dementia requires institutionalisation and care both day and night. In the US alone the cost of dementia is estimated to be US$1 trillion [30]. Postponing the onset of severe dementia that requires institutionalisation one year is calculated to save $48,096 (€43,259) per patient [2].

There are studies that seem to imply that an early diagnosis of dementia can reduce these costs for society; however, it is hard to definitively prove that it is the case. Weimer et.al. claims that an early diagnose, defined as having a mini mental state examination (MMSE) number at 28 or above, can save over $10,000 per diagnosis in the United States [29]. Weimer et.al. reaches this number by assuming that drugs can cut the rate of MMSE points decline by half which in turn reduces the years spent in nursing homes by 1.2 years according to their Monte Carlo simulation. Since the state pays for the care in the nursing homes, this cost reduction would directly lead to saved tax money [29]. However, we see problems with Weimer et.al.'s paper. The assumption that some drugs can lower the reduction rate of MMSE points can be strongly questioned. This assumption is based on one study [18] which achieves these results; yet, a widely cited and large meta study on the same category of drugs found no such meaningful reduction [23]. If the assumption of the effectiveness of drugs falls, then the entire reasoning around the cost reductions falls as well. Still, there are other studies that show evidence that early diagnostics lead to cost reduction for the state. Mittelman et.al. has a study which shows that counseling for spouses who care for patient with dementia can reduce the time to institutionalization by 1.5 years [20]. This counseling becomes more effective if applied early according to [18]. Thus, the state could potentially save substantial amounts by investing in screening tools that diagnose dementia early.

C. Social factors

Some believe that getting a diagnosis is a human right. This belief is reinforced by surveys where subjects without MCI or dementia were asked whether they would like to be screened for dementia. 98% of participants in the survey answered that they would want to be screened and 99% were willing to take medication if the medication would reduce the risk for getting dementia [8]. It is clear that most would prefer to know they have dementia as soon as possible. It is possible that the idea that a diagnosis is a human right gains political traction. In such a scenario, our app for early dementia screen would have greater market potential since there would be a higher willingness to pay.

D. Technological factors

New technologies within computer science, AI and machine learning have in recent years greatly contributed to new and effective methods. Indeed, as both the method described in this paper and the one by Mindmore have shown, there is great potential for automated methods of diagnosis relying on apps and machine learning.

One technological limitation of our method is the need for data. If the method is to be effective at scale it would need thousands of recordings from dementia patients. This data is needed for every natural language which makes it harder to apply the method in new countries. However, there is a lot of scientific research in this area and many research teams are collecting data. It would be possible to collect this data from various universities across the world, and thus, this technical limitation is not a big one.

VII. SWOT

In order to supplement the market analysis provided by Porter's five forces and PEST, we analyze the market with the SWOT framework. We evaluate the remaining strengths, weaknesses, opportunities and threats.

A. Strengths

There are multiple factors that strengthen the market potential of an app-based early screening of dementia using natural language processing. First of all, it is a product which can be used in countries and regions which have low resources. Most of the undiagnosed cases today are in areas of low socioeconomic status. According World Alzheimer Report, up to 90% of cases in areas with low socioeconomic status can be undiagnosed [22]. The product could potentially help millions of people who would remain undiagnosed without our help. While that probably would have limited economic impact directly, it could give the app strong political support or sponsors. The cost of using the app once developed is very low. Even though the machine learning algorithm requires the use of a GPU, each diagnosis will cost less than one dollar in terms of the marginal cost. Therefore, we could offer these services pro bono to areas that lack the required resources for a paid version. Furthermore, the method performs very well compared to many other machine learning methods, and could potentially provide a more objective judge than a human counterpart.

B. Weaknesses

One of the key factors that affect the cost savings with an early diagnose of dementia is whether the preemptive measures are likely to postpone institutionalisation. World Alzheimer Report screened 8039 papers for quantitative evidence that early preemptive measures can postpone institutionalisation [22]. Unfortunately, they could not find any considerable quantitative evidence that clearly shows that early preemptive measure can reduce the risk of institutionalisation. It is very difficult to conduct research that show reliable evidence that early preemptive measures can postpone institutionalisation since such research would have to been done over the course of over twenty years, according to our interview with the health care professional expert. However, such research would to be provided in order for governments to make the investments needed in a dementia screening tool. According to our expert interview, there is research currently in progress in this field which is showing promising results, it will be published in the coming months. Hopefully such research shows that an early diagnose could postpone institutionalisation.

Both our interviewees mentioned the importance of rigorous scientific research when convincing government agencies such as Socialstyrelsen. Here our method is lacking since it it still in its infant stages. A large research study with hundreds of patients over the course of years would have to be conducted in order to get the required scientific support for our method.

Furthermore, our method currently only provides the potential to automate one part of the diagnostic tests. While certainly useful, it would help to include automation for various other tests in our product.

C. Opportunities

A big opportunity is the large size of the market for dementia diagnostics and the possibility to scale quickly worldwide. It is estimated that there are 28 million people with dementia who are undiagnosed [22]. Furthermore, the number of undiagnosed patients will continuously increase with the rate of the ageing population. If the cost of one diagnose is $100, the total market for dementia diagnostics could be estimated to $2,8 billion. While this estimate is a simple one, it says something about the sheer size of the market.

There is currently no single firm that dominates the market for digital dementia screening tools. This implies that the sector currently does not have a dominant design. If one company can take a big market share quickly, there is the possibility to create the dominant design. Getting to a dominant design has considerable benefits such as brand loyalty, giving customers high switching costs, access to scarce resources such as talent and much more [26].

D. Threats

There is a risk that the proliferation of dementia screening tools might lead to over diagnosis. With some screening tools there may be a risk of diagnosing mild cognitive impairments as dementia. Only 5-10% of people with mild cognitive impairment will progress to dementia each year, and as many as 40-70% of people do not progress or their cognitive function may even improve [16]. It is extremely important for the screening tool to lower the amount of false positives since that might lead to over diagnosis. However, this is a challenge since it might be even more important to avoid false negatives, since that could have devastating consequences for a patient that does not receive care. This issue is one of the biggest challenges related to dementia screening tools.

VIII. Conclusion

First, we note that our primary source sample is small and as such it is difficult to draw strong conclusions. However,

if the sample is representative, we can state the following things. Dementia is a massive cost to society today, both economically and socially. Postponing the onset of severe dementia and the subsequent institutionalisation by one year is estimated to save $48,096 per patient. There are some evidence that preemptive measures can postpone institutionalisation by several years, and these are most effective if applied early. Therefore, the market potential of an effective screening tool for dementia should be millions of dollars, since it could save billions for society if it can postpone institutionalisation. However, the market potential still contains some uncertainty due to (1) the uncertain effectiveness of preemptive measure for reducing institutionalisation and (2) uncertainty regarding just how early the screening tools can detect dementia. Despite these uncertainties there is considerable interest from states in automatic dementia screening tools which means that the market potential should be big.

IX. Author contributions

Alfred Wahlforss BSc student in industrial engineering and management at KTH. Incoming MSc student in data science at Harvard University. Main focus in this study was data analysis, data engineering, BERT modelling, literature study and more.

Alexander Aslaksen Jonasson MSc student in industrial engineering and management at KTH, specializing in Machine Learning. Main focus in this study was data analysis, data engineering, machine learning modelling, literature study and more.

References

[1] Transformer transformer: A novel neural network architecture for language understanding. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html. Accessed: 2020-05-01.

[2] Örjan Åkerborg, Andrea Lang, Anders Wimo, Anders Sköldunger, Laura Fratiglioni, Maren Gaudig, and Mats Rosenlund. Cost of dementia and its correlation with dependence. *Journal of aging and health*, 28(8):1448–1464, 2016.

[3] Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228, 2017.

[4] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

[5] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

[6] Joan C Borod, Harold Goodglass, and Edith Kaplan. Normative data on the boston diagnostic aphasia examination, parietal lobe battery, and the boston naming test. *Journal of Clinical and Experimental Neuropsychology*, 2(3):209–215, 1980.

[7] Lawrence P Carr and Alfred J Nanni Jr. *Delivering Results: Managing What Matters*. Springer US, New York, NY, 1 edition, 2009.

[8] William Dale, Gavin W Hougham, Emily Kay Hill, and Greg A Sachs. High interest in screening and treatment for mild cognitive impairment in older adults: A pilot study. *Journal of the American Geriatrics Society*, 54(9):1388–1394, 2006.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] V Olga B Emery. Language impairment in dementia of the alzheimer type: A hierarchical decline? *The International Journal of Psychiatry in Medicine*, 30(2):145–164, 2000.

[11] Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. Detecting alzheimer's disease from continuous speech using language models. *Journal of Alzheimer's Disease*, 70(4):1163–1174, 2019.

[12] Marilyn M Helms and Judy Nixon. Exploring swot analysis–where are we now? *Journal of strategy and management*, 2010.

[13] Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268, 2018.

[14] Fakultit Informatik, Y. Bengio, Paolo Frasconi, and Jfirgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*, 03 2003.

[15] Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):9, 2015.

[16] David G Le Couteur, Jenny Doust, Helen Creasey, and Carol Brayne. Political drive to screen for pre-dementia: not evidence based and ignores the harms of diagnosis. *Bmj*, 347:f5125, 2013.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[18] Oscar L Lopez, James T Becker, Judith Saxton, Robert A Sweet, William Klunk, and Steven T DeKosky. Alteration of a clinically meaningful outcome in the natural history of alzheimer's disease by cholinesterase inhibition. *Journal of the American Geriatrics Society*, 53(1):83–87, 2005.

[19] Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017*, pages 232–237, 2017.

[20] Mary S Mittelman, William E Haley, Olivio J Clay, and David L Roth. Improving caregiver well-being delays nursing home placement of patients with alzheimer disease. *Neurology*, 67(9):1592–1599, 2006.

[21] Michael Porter. The five competitive forces that shape strategy. *Harvard Business Review*, 86(1):78–93, 2008.

[22] M Price, C Bryce, and C Ferri. World alzheimer report 2011. *Alzheimer's Disease International, London*, 2011.

[23] Parminder Raina, Pasqualina Santaguida, Afisi Ismaila, Christopher Patterson, David Cowan, Mitchell Levine, Lynda Booker, and Mark Oremus. Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. *Annals of internal medicine*, 148(5):379–397, 2008.

[24] Alexander Rush. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[25] Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195, 2015.

[26] Linda F Tegarden, Donald E Hatfield, and Ann E Echols. Doomed from the start: What is the value of selecting a future dominant design? *Strategic Management Journal*, 20(6):495–518, 1999.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[28] Sebastian Wankerl, Elmar Nöth, and Stefan Evert. An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language. In *INTERSPEECH*, pages 3162–3166, 2017.

[29] David L Weimer and Mark A Sager. Early identification and treatment of alzheimer's disease: social and fiscal outcomes. *Alzheimer's & Dementia*, 5(3):215–226, 2009.

[30] Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, A Matthew Prina, Bengt Winblad, Linus Jönsson, Zhaorui Liu, and Martin Prince. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia*, 13(1):1–7, 2017.

[31] Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. Speech recognition in alzheimer's disease and in its assessment. In *Interspeech*, pages 1948–1952, 2016.

TRITA -EECS-EX-2020:239