Analysing Representations of Memory Impairment in a Clinical Notes Classification Model

Mark Ormerod, Jesús Martínez del Rincón, Neil Robertson
Bernadette McGuinness, Barry Devereux
Queen's University Belfast
{*mormerod01, j.martinez-del-rincon, n.robertson, b.mcguinness, b.devereux*}*@qub.ac.uk*

Abstract

Despite recent advances in the application of deep neural networks to various kinds of medical data, extracting information from unstructured textual sources remains a challenging task. The challenges of training and interpreting document classification models are amplified when dealing with small and highly technical datasets, as are common in the clinical domain. Using a dataset of de-identified clinical letters gathered at a memory clinic, we construct several recurrent neural network models for letter classification, and evaluate them on their ability to build meaningful representations of the documents and predict patients' diagnoses. Additionally, we probe sentence embedding models in order to build a human-interpretable representation of the neural network's features, using a simple and intuitive technique based on perturbative approaches to sentence importance. In addition to showing which sentences in a document are most informative about the patient's condition, this method reveals the types of sentences that lead the model to make incorrect diagnoses. Furthermore, we identify clusters of sentences in the embedding space that correlate strongly with importance scores for each clinical diagnosis class.

1 Introduction

While the majority of clinical data is made up of structured information (Jee and Kim, 2013), which can often be readily integrated into data models for research, there is a significant amount of semi-structured and unstructured data which is increasingly being targeted by machine learning practitioners for analysis. As a general rule, this unstructured data is more difficult to analyse due to an absence of a standardised data model (Ann Alexander and Wang, 2018). Unstructured clinical data includes a variety of media, such as video, audio, image and text-based data, with the majority of such data being made up of text and images. Recently, there has been a series of breakthroughs in the application of machine learning techniques for medical imaging data in order to achieve expert-level performance on diagnosis tasks (Rajpurkar et al., 2017). However, machine learning models using semi-structured and unstructured textual data from the clinical domain have received less attention and to date have not seen the same degree of successful application. Examples of unstructured medical data featuring "free text" include discharge summaries, nursing reports and progress notes. Historically, one of the challenges of applying natural language processing (NLP) methods to clinical data has been the often limited amount of data available, which has traditionally necessitated a reliance on manual feature engineering and relatively shallow textual features (Shickel et al., 2018).

Taking a novel dataset of labelled clinical letters compiled at a memory clinic as the target data domain, we build state-of-the-art deep learning models for the task of clinical text classification, and evaluate them on their ability to predict a clinician's diagnosis of the patient. However, deep learning models generally require very large training datasets. Our approach to the problem therefore incorporates transfer learning, and we make use of embedding data from pre-trained models trained on large corpora. In order to investigate the relative usefulness of word-level and sentence-level information, we train and evaluate several models, including a ULMFiT model (Howard and Ruder, 2018) and two long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) models: one trained on word embedding representations of the documents and one trained on sentence embedding representations (Basile et al., 2012).

An infamous problem of deep neural networks is that they are "black boxes", with the details of how they represent and process information

Proceedings of the BioNLP 2019 workshop, pages 48–57
Florence, Italy, August 1, 2019.

being uninterpretable to humans. To shed light on how a recurrent neural network models clinical documents in order to correctly predict a patient's diagnosis, we investigate two complementary approaches to model interpretation. Firstly, we develop a simple measure of sentence importance and demonstrate its effectiveness in interpreting a complex LSTM model's decision making process. Secondly, we discover clusters in the high-dimensional space of the sentence embedding model and test their correlation with feature importance scores for a given diagnosis class. This analysis yields insights into a model's representation of the clinical notes, allowing us to automatically extract clusters of sentences that are most relevant to the model's predictions.

2 Related Work

Document classification is a well-researched task in NLP that has been tackled using a wide variety of machine learning models, such as support-vector machines (SVMs) (Manevitz and Yousef, 2001), convolutional neural networks (CNNs) (Conneau et al., 2016) and recurrent neural networks (RNNs) (Yogatama et al., 2017). In the clinical domain, document classification models have been used in diverse tasks such as predicting cancer stage information in clinical records (Yim et al., 2017), extracting patient smoker-status from health records (Wang et al., 2019) and classifying radiology reports by their ICD-9CM code (Garla and Brandt, 2013). The problem of categorising clinical free text documents is closely related to several subtasks in the area of Electronic Health Record (EHR) analysis, including information extraction and representation learning. Information extraction is an umbrella term that covers diverse subtasks such as expanding abbreviations using contextual information, and the automatic annotation of temporal events (e.g. mapping from inputs such as "The patient was given stress dose steroids prior to his surgery" to output "[stress dose steroids] BEFORE [his surgery]" (Sun et al., 2013). Other NLP problems in this field that are relevant to free text analysis are outcome prediction and de-identification.

There are many ways to construct a representation of the input data that can be provided to a document classification model. A popular alternative to older approaches to text representations, such as bag-of-words (BoW), is to embed the input tokens in a high-dimensional vector space, resulting in each word being mapped to a list of real-valued numbers (a "word embedding"). One simple method of extracting word embeddings involves concatenating the hidden layer activations observed in a trained language model after processing all words up to the target word. As language models automatically learn rich semantic and syntactic features of words, these embeddings can provide valuable input features for downstream information extraction tasks. While the dimensions in the embedding space can correspond to interpretable features, this is not generally the case. However, a major motivation for using word embeddings is the ability to re-use pre-trained embeddings, essentially resulting in a form of transfer learning (Pan et al., 2010). In this study we use 300-dimensional fastText word embeddings (Bojanowski et al., 2017) which were pre-trained on the Common Crawl dataset using the skipgram schema (Mikolov et al., 2013), which involves predicting a target word based on nearby words.

Class	# doc.	# sent.	# sent. (masked)
D	32	1420	1225
M	30	1140	985
N	44	1767	1547

Table 1: Number of documents and sentences in the clinical notes dataset. *D*: Dementia, *M*: MCI, *N*: Non-impaired.

Similar to word embeddings, sentence embeddings are high-dimensional vectors that can represent features of a sequence of words. Our use of sentence embeddings is motivated by the fact that, for small amounts of data, it may be more difficult for a recurrent neural network to capture diagnosis-relevant dependencies over many word vectors than it is to classify a document made up of a smaller number of semantically richer sentence vectors. In this study we use 4096-dimensional InferSent embeddings (Conneau et al., 2017) that were extracted from a model pre-trained on the Common Crawl dataset.

After training recurrent models using these state-of-art NLP techniques to predict the diagnosis class associated with each document, we explore ways of visualizing and understanding how the models incorporate these vectors in order to make accurate predictions.

Model	Accuracy	Precision	Recall	F1 Score
Random	0.333	0.333	0.333	0.333
Max. class	0.415	0.138	0.333	0.196
BoW+Random Forest	0.425	0.417	0.413	0.414
LSTM (*fastText word emb.*)	0.543	0.636	0.502	0.502
LSTM (*InferSent sentence emb.*)	**0.690**	**0.702**	**0.669**	**0.674**
ULMFiT	0.571	0.437	0.500	0.440

Table 2: Results (average over 5 folds) for the diagnosis classification task for the masked dataset. Precision, recall and F1 score are macro-averaged across the classes.

3 Data

We collected a corpus of consultation reports compiled by clinicians at a memory clinic to use as the data domain for the document classification task. Each report is anonymised and describes the clinician's review of a patient who suffers from memory or cognitive issues. Each report is labelled by one of three classes, corresponding to the diagnoses of *dementia*, *mild cognitive impairment (MCI)* and *non-impaired*. The documents can be considered semi-structured, as they are made up of free-text details that follow a loose narrative trajectory. The notes typically begin with a description of the patient's history and symptoms, and ultimately conclude with recommendations on how to proceed which may include scheduling a follow-up appointment, arranging further tests, or organising a treatment course based on the available evidence.

From this corpus, we build a version of the notes in which explicit diagnostic information is masked out. For example, the sentence *"We would recommend commencing on a Rivastigmine patch 4.6 mg for 24 hours and then to be increased to 9.5 mg for 24 hours once daily if tolerated."* would not be included in the masked diagnosis dataset, as the drug Rivastigmine is used to treat mild to moderate Alzheimer's disease and Parkinson's, and so its mention here trivially identifies the diagnosis. In this work, we are interested in the ability to make predictions from more subtle diagnostic signals, requiring our model to build semantic representations of cognitive impairment that go beyond counting the occurrence of single words. Table 1 presents summary metrics of the datasets.

Deep learning models are generally trained and tested on very large datasets, in contrast to the small corpus of demential letters that we have gathered, and in contrast to clinical note databases generally. This motivates our use of transfer learning.

Tackling the problems of training and interpreting models trained on datasets of this scale is directly relevant to the real world challenges of using natural language processing to support clinical decisions, such as identifying patients who may be applicable to participate in a clinical trial (Sarmiento and Dernoncourt, 2016). Annotating gold-standard training examples for such problems is resource intensive (Savkov et al., 2016). We would therefore like to build robust and general models given a small amount of samples. Recent work on training large language models on massive amounts of data thus has much potential for zero-shot classification of natural language documents (Yogatama et al., 2017).

4 Models and Evaluation

We investigate the relative performance of LSTM models trained with a sequence of word embeddings, LSTM models trained with a sequence of sentence embeddings, and a state-of-the-art document classification model, ULMFiT. One motivation for choosing these experimental models is to investigate which models can capture long-term dependencies across a clinical document, given a relatively small amount of samples (*n=106*). In addition to these three models, we also test a random forest baseline model, a model that randomly selects the class and a model that chooses the most common class (which is *non-impaired*). The random forest model is trained to classify a document based on its bag-of-words representation. All models are cross-validated using 5 folds of

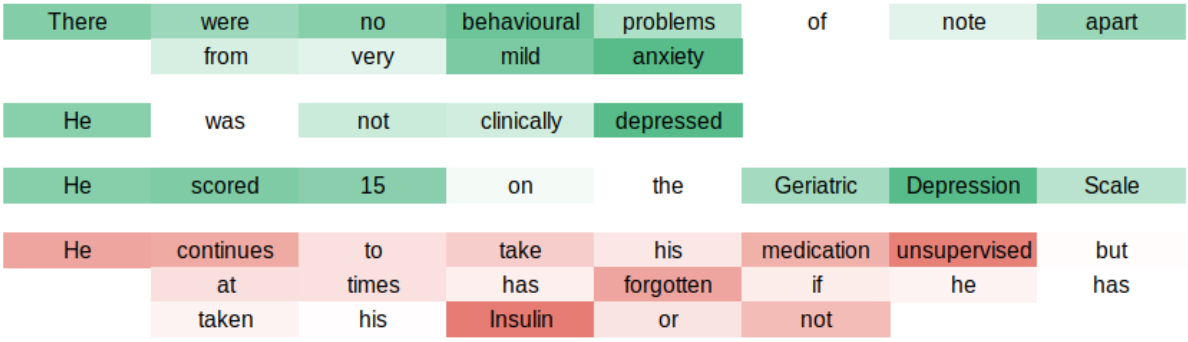


Figure 1: Visualisation of sentence importance with respect to the successful classification of *non-impaired* for a subset of a document. Sentences that were found to be important for the classification of *non-impaired* are coloured green while a sentence that increases the chance of a misclassification (i.e. an incorrect *MCI* diagnosis) is coloured red. The saturation of the colours corresponds to how much a given word contributes to a sentence's InferSent embedding

the dataset, ensuring that the class distribution is equal across all folds. The ULMFiT model is pre-trained on the Wikitext-103 dataset (Merity et al., 2017) and fine-tuned using default hyperparameters (*fine-tuning epochs=25, fine-tuning batch size=8, fine-tuning learning rate=0.004, training epochs=50, training batch size=32, training learning rate=0.01*) which have been shown to be robust across various tasks (Howard and Ruder, 2018). The LSTM model's hyperparameters were chosen by a grid-search. Both the sentence embedding LSTM and the word embedding LSTM were made up of one hidden layer with 256 hidden units.

The classification results for the models for the masked dataset are presented in Table 2. Each of our three models perform significantly better than chance and better than the random forest baseline model, with the LSTM model trained with sentence-embedding sequential input achieving the best performance. For this amount of training data, we would expect models that are trained on shorter sequences of more semantically enriched pre-trained vectors (i.e. sentence embeddings) to perform better than much longer sequences of vectors with less dimensions (i.e. word embeddings). This is because much of the work of combining word-level tokens into a contextual representation that is relevant to a statistical model of human language has already been done when training with pretrained representations extracted at the sentence-level. Somewhat surprisingly, the model trained on sentence embeddings outperformed the fine-tuned ULMFiT. Future work may shed light on how the amount of training samples can affect the choice of whether to use fine-tuning or pre-trained embedding representations as model input.

5 Model Interpretability: Calculating Sentence Importance Scores

After demonstrating the effectiveness of using pre-trained sentence embeddings to classify the clinical documents, we investigated model interpretability by calculating a measure of the importance of each sentence in the sequence of sentences to the model's prediction for a document. We propose a measure of feature importance based on perturbative approaches to variable importance (Breiman, 2001), which estimate the importance of variables by iteratively randomly perturbing each variable and observing the change in loss. This technique is similar to measuring information gain (Quinlan, 1986), but rather than selecting important components of fixed input, we rate the importance of a sentence vector in the sequence of sentence vectors presented to our sequential LSTM classifier. For example, in order to generate the importance score for the first sentence in a document made up of *m* sentence embeddings, we construct an augmented version of the document containing all but the first sentence, and examine the resulting change in the prediction for that document. More formally, for sentence *n*, we generate the following version of the document d (with ground truth label c) with sentence n removed:

$$d_n = [s_0, s_1, \ldots, s_{n-1}, s_{n+1}, \ldots, s_{m-1}]$$

Next, the augmented document d_n is fed into the trained LSTM (using the best-in-fold model

Ratio	Sentence
-3.469	"He and his wife both report agitation disinhibition and irritability"
0.078	"He would say that he feels depressed at times"
0.149	"She was tremulous which <NAME> felt was most likely due to anxiety"
	. . .
2.108	"He had an equivalent score of 19 / 30 on the MMSE"
8.105	"He had an equivalent score of 29 / 30 on the MMSE"
12.887	"He had an equivalent score of 22 / 30 on the MMSE"

Table 3: Sentences sorted by feature importance for a correct diagnosis of *non-impaired*. Sentences with low scores do not support a prediction of *non-impaired* within the context of the corresponding clinical letter.

from Section 3, which achieved an accuracy of 73%) and we measure the network's output logit for the correct class. The importance score is calculated as the ratio of the model's output for the correct class excluding the sentence to the model's output for the correct class including a given sentence.

$$ratio_n = \frac{logit(c \mid d_n)}{logit(c \mid d)}$$

The most important sentences minimise this ratio. When the ratio is over 1, the inclusion of the sentence in the document leads to a smaller probability of selecting the correct class, and so sentences that maximise the ratio are the most misleading sentences with respect to the correct classification. Examples of highly important and highly misleading sentences across the corpus for a diagnosis of *non-impaired* are presented in Table 3. The average sentence importance trajectory over each class was also investigated and is presented in in Figure 2.

Figure 1 presents a section of a clinical letter for a patient with a diagnosis of *non-impaired*, with sentences coloured green or red depending on whether they increase or decrease the chance of correctly classifying the document. Within each sentence, the contribution of a word to the InferSent sentence embedding is visualised by colour saturation. We can see that the importance measure provides intuitive insights into how the recurrent neural network models the document. For example, the final sentence in Figure 1 decreases the chance of classifying the document as *non-impaired* because it states that the patient sometimes forgets to take their medicine – in isolation this sentence could naively be considered to imply a diagnosis of memory impairment, but as the model processes the full document it is able to

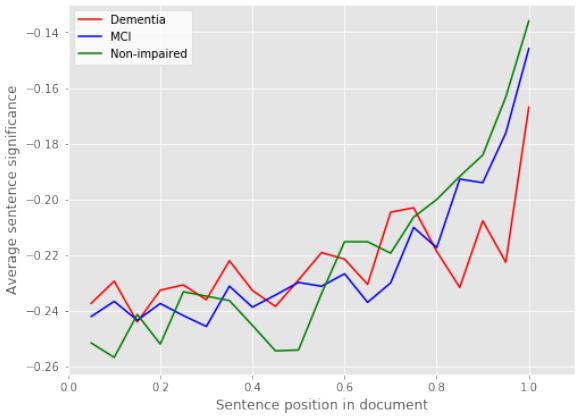
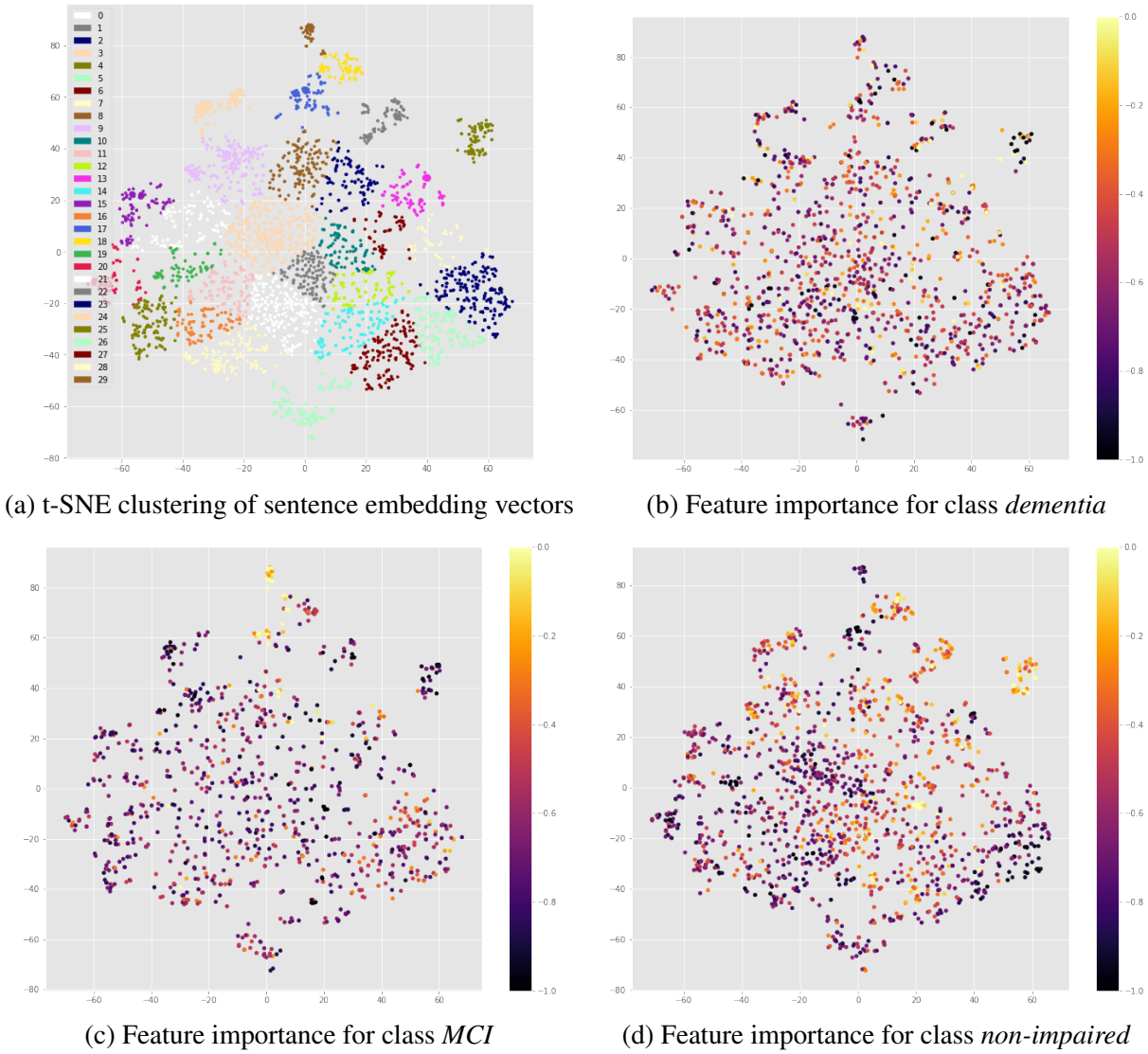


Figure 2: Average sentence importance over each class, as a function of sentences' position in the texts. Sentence importance ratios are normalised within each document and split by in-document position into 20 bins. For each class, we plot the negative of the average for each bin.

accumulate evidence and predict the correct diagnosis. By examining the contribution of each word to the InferSent vectors, we can see that negating words such as "not" are handled appropriately within the sentence embedding (e.g. "not clinically depressed" increases the probability of a correct *non-impaired* classification). Our model interpretation technique therefore demonstrates how the LSTM sentence embedding model improves on the simple bag-of-words baseline, where the word "depressed" would be incorrectly taken as negative evidence for a non-impaired diagnosis.

6 Cluster Analysis

In order to investigate the relationship between sentence importance and the sentence embedding space, we performed a cluster analysis. The 4096-dimensional sentence embeddings were projected



(a) t-SNE clustering of sentence embedding vectors

(b) Feature importance for class *dementia*

(c) Feature importance for class *MCI*

(d) Feature importance for class *non-impaired*

Figure 3: 2-dimensional projection of sentence embedding vectors. **(a)**: 30 clusters were identified and labelled using mean shift clustering. **(b) - (d)**: Heat maps of sentence vectors coloured by sentence importance for each class reveal clusters of sentences that are relevant to a given diagnosis. Colour scales indicate normalised values; brighter colours indicate more important sentences.

to two dimensions using t-SNE (van der Maaten and Hinton, 2008). We used the mean shift clustering technique (Yizong Cheng, 1995), an algorithm that does not require the number of clusters to be specified in advance, to discover clusters of similarly represented sentences in this space (Fig. 3(a)). Sentences that are important for the model's classification of a specific diagnosis are visualised by colouring the sentences using the corresponding importance score. This step was performed for each of the three classes (Fig. 3(b)-(d)).

Correlation tests were used to investigate the relationship between sentence clusters and their importance to a model's prediction for each class. For each class c and for each cluster cl, we first gather the sentences that appear in documents of class c. Next, we assign each sentence a value of 1 or 0 depending on whether the sentence is in cluster cl. Using Spearman's Rho, we calculate the correlation between this value and the sentences' importance scores for the given class. In each trial, sentences that do not appear in documents of the target class are excluded. The results reported in Table 4 show the clusters that were found to be significantly correlated with at least one of the classes' importance scores. It was found that 15 out of the 30 automatically discovered sentence clusters can be considered significantly important

in the model's decision making.

To assist in interpreting the information captured by each cluster, we depict the clusters using the most frequent bigrams across all of that cluster's sentences (Table 4). For example, one cluster (corresponding to cluster 20 in Figure 3(a)) contains sentences that mention the individual's family (significantly positively associated with a *non-impaired* diagnosis), while cluster 22 corresponds to sentences about the patient's blood pressure and heart rate (significantly negatively associated with a *non-impaired* diagnosis). Again, these results show the utility of combining sentence importance measures with sentence embeddings to reveal the clinically relevant detail in the documents.

7 Discussion

The results presented in Table 3 demonstrate the sentences that are most significant and most misleading for the LSTM InferSent model with respect to the diagnosis of non-impairment. We can see that the most significant sentences are those that refer to patients' mood and anxiety disorders. These types of sentences are over-represented in the *non-impaired* group. The types of sentences that are most misleading to the diagnosis of *non-impaired* are those of the format "*[pronoun] had an equivalent score of [score] / 30 on the MMSE*". An obvious question regarding this result is whether information about MMSE scores can be represented by the InferSent embeddings in such a way as to distinguish it from other sentences that differ only, but importantly, by a single integer value. We can see that the relationship between the significance of the sentence to the actual results in the sentence is non-linear. The 84 mentions of the Mini-Mental State Examination (MMSE) test are equally divided across the 3 classes; as there are more *non-impaired* documents in the dataset overall, the model benefits from learning not to predict this diagnosis when it encounters any sentence embedding in the MMSE cluster (cluster 17 in Figure 3(a); the corresponding points in Figure 3(d) indicate their decreased importance for this category). Further analysis may include using *diagnostic classifiers* (Hupkes et al., 2018) to test whether a model can accurately decide whether the first of two given sentence embeddings reports a larger score.

Figure 2 shows the average sentence significance across the documents for each of the three classes. For all classes, we can see that the importance of sentences tends to increase with their in-document position. This trend may correspond to the semi-structured nature of the documents, reflecting information becoming more relevant to a diagnosis towards the end of a document. Another possible explanation could be that the recurrent neural network is unable to capture long-distance dependencies given the small amount of samples in the dataset, resulting in a kind of recency bias in the model's processing (since the model only makes its prediction at the end of the sequence of sentences). Further work may involve systematically changing the position of each sentence within each document in order to investigate the effect that this has on the importance scores associated with each sentence.

Table 4 shows that no clusters were significantly correlated with the class *dementia*, with all reported clusters being significantly correlated with at least one of *MCI* or *non-impaired*. Excluding cluster 18, all of the clusters that are significant for both *MCI* and *non-impaired* form pairs of negative vs. positive correlations between these two classes, suggesting that the model learns primarily to discriminate between these classes. Examining the confusion matrix for the model, we found that the model has a true positive rate of 1.0 and 0.89 for *MCI* and *non-impaired*, and minimises the amount of false positives between these two classes. However, the model performs poorly when the actual document corresponds to a diagnosis of *dementia* (with a true positive rate of 0.29). This is consistent with the observation that none of the clusters significantly correlate with this class. While this insight could be gained from examining the confusion matrix alone, the advantage of employing the interpretation methods developed in this paper is that they allow us to gain an understanding of how the model's processing of sentences over time leads to these inequalities, suggesting avenues of attack for constructing more accurate representations of the documents going forward.

In future work, we plan to gather more clinical documents that describe patients with memory impairment and continue our analysis of language modelling and classification in this distribution. We hope to subsequently apply state of the art contextualised embeddings such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018)

Cluster	Top bigrams in cluster	Rho_D	Rho_M	Rho_N
2	"behavioural problems", "neurological deficit", "extra pyramidal"	0.036	**–0.215*****	0.142***
3	"short term", "years ago", "poor short"	–0.022	0.032	**0.119*****
5	"family history", "disease dementia", "alzheimers disease"	0.039	**–0.147*****	0.146***
7	"activities daily", "daily living", "remains independent"	–0.013	–0.121*	**0.155*****
9	"medical history", "ischaemic heart", "heart disease"	–0.003	**0.161*****	–0.131***
10	"memory fluency", "verbal fluency", "points lost"	–0.033	**0.133****	–0.097*
12	"misplacing items", "cognitive checklist", "disorientation time"	0.000	**0.143*****	–0.098*
17	"30 mmse", "mmse equivalent", "29 30"	–0.040	**–0.193*****	0.112***
18	"cognitive testing", "100 ace", "addenbrooke cognitive"	–0.010	–0.171***	**–0.181*****
20	"unaccompanied morning", "four children", "two children"	0.022	–0.029	**0.165*****
22	"blood pressure', "bpm regular", "examination pulse"	–0.045	0.130**	**–0.124*****
23	"b12 folate", "screening bloods", "thyroid function"	–0.022	0.022	**–0.089***
24	"current medications", "mg daily", "40 mg"	0.021	**0.181*****	–0.106**
25	"geriatric depression", "depression scale", "scored 15"	0.010	**0.182*****	–0.229
27	"onset progression", "progression described", "physical examination"	–0.064	0.132**	**–0.113*****

Table 4: Automatically discovered sentence clusters that significantly correlate with sentence importance for at least one class. For each cluster and for each class, we use Spearman's Rho to test the correlation between a sentence's importance with respect to the class of interest, and whether or not the sentence is in the given cluster. The most frequent within-cluster bigrams were extracted after removing stop words from the sentences. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, Bonferroni corrected. *D*: Dementia, *M*: MCI, *N*: Non-impaired.

to a larger corpus in order to further use feature extraction to build and understand meaningful semantic representations of cognitive impairment as described by clinicians. As part of this work, we aim to examine how models trained on the writing style of one clinician apply to those written by others, as the corpus used in this study was sourced from a small number of clinicians. We suspect that analysing a model's inter- and intra-clinician performance metrics will yield useful insights into how well the model has generalised, and how clinicians may differ in terms of the subtle but diagnosis-relevant information they include in the documents.

8 Conclusion

We showed the effectiveness of using pre-trained sentence embeddings and recurrent neural networks for a document classification task using a corpus of natural language clinical reports. The sentence-level LSTM model performed better than both an LSTM trained on word embeddings and a simple bag-of-words baseline. Following this result, we developed a simple and intuitive perturbative measure of sentence importance for the sentences in the corpus. After demonstrating how this measure can be used to interpret the success and failure cases of a trained model, we used cluster analysis to identify regions in the sentence embedding space that are significantly correlated with sentence importance for specific diagnosis classes.

By reviewing the most frequent bigrams in each cluster and examining the sign of Spearman's Rho for each corresponding correlated class, we can interpret how differential processing of sentence vectors within each cluster can lead to class imbalances in the model's predictions, demonstrating the power of our approach for model interpretability and evaluation.

Acknowledgements

We would like to thank the three anonymous reviewers, Stuart Millar, and Steven Derby for their feedback and suggestions. This work was part-funded by a Data Analytics Dementia Pathfinder Programme Grant from the Northern Ireland HSCB eHealth Directorate.

References

Cheryl Ann Alexander and Lidong Wang. 2018. Big Data and Data-Driven Healthcare Systems. *Journal of Business and Management Sciences*, 6(3):104–111.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2012. A Study on Compositional Semantics of Words in Distributional Spaces. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 154–161. IEEE.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

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

Vijay N Garla and Cynthia Brandt. 2013. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association*, 20(5):882–886.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:328–339.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*.

Kyoungyoung Jee and Gang-Hoon Kim. 2013. Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System. *Healthcare Informatics Research*, 19(2):79.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Larry M Manevitz and Malik Yousef. 2001. One-class SVMs for document classification. *Journal of machine Learning research*, 2(Dec):139–154.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *preprint arXiv:1301.3781*.

Sinno Jialin Pan, Qiang Yang, and Others. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

Raymond Francis Sarmiento and Franck Dernoncourt. 2016. Improving patient cohort identification using natural language processing. In *Secondary analysis of electronic health records*, pages 405–417. Springer.

Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language resources and evaluation*, 50(3):523–548.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):814–9.

Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J. Atkinson, Shreyasee Amin, and Hongfang Liu. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 19(1):1.

Wen-Wai Yim, Sharon W Kwan, Guy Johnson, and Meliha Yetisgen. 2017. Classification of hepatocellular carcinoma stages from free-text clinical and radiology reports. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:1858–1867.

Yizong Cheng. 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799.

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*.