Unsupervised learning over unstructured clinical data for identifying patterns and incidental findings

Part I: Methodology

Dataset

The dataset, consisting of 6,117 free-text reports on magnetic resonance imaging (MRI) scans of the cranial cavity of patients examined for vestibular schwannoma was provided by the Royal Berkshire Hospital, UK. The reports were first automatically anonymised and then manually labelled by a clinician, as those mentioning tumour (positive class) and not (negative class).

Technology selection

Google Colab was the primary platform used to undertake the experiments outlined in this report. The online capabilities allowed switching between Python 3 Google Compute Engine backend GPU and TPU, based on the runtime needs of the models. In addition, 35.24 GB of RAM was allocated, making the computing resources more than adequate for the intended purpose. The software selection was based on the domain of natural language processing (NLP), a specific subdomain in linguistic and artificial intelligence. All software selected was open source and these libraries include and are not limited to Python (v 3.7), Tensorflow (v 2.6), Sentence Transformers (v 2.2.2), Gensim (v 3.6.0), NLTK (v 3.7) and NLPAUG (v 1.1.11).

Exploratory data analysis

Exploratory data analysis was undertaken first to better understand the data and the problem scope. Figure 1 and Figure 2 show samples of reports classified as negative (i.e. no tumour found) and positive (a tumour is found), respectively. There were 5,882 negative and 235 positive reports, meaning that the dataset was highly imbalanced, which is typically the case in the healthcare domain. The study considered data augmentation to address the class imbalance, which is known to degrade performance of machine learning (ML) models.

```
MR202021468 25/11/2020 MRI Internal auditory meatus Both

Clinical History:
Exclude vestibular schwannoma. Lt asymm HL. Patient has MS. Feels she would
requiremedication to undertake MRI


Findings:
Both internal auditory meati and membranous labyrinths appear normal.
No cerebellopontine angle lesion or posterior fossa abnormality is seen. In particular,
no vestibular schwannoma is identified.

Tortuosity of the left vertebral artery is incidentally noted which is seen indenting
and mildly displacing the brainstem to the right side. This is of no further clinical
relevance.


Dr ##### #####
Consultant Radiologist
GMC #######
```

Figure 1: *A radiology report sample of a negative tumour outcome.*

```
MR172001711 03/11/2017 MRI IAM with contrast Both

Clinical History:
known 5mm left acoustic neuroma. serial scan. November 2017 please


Findings:
Comparison with images from January 2017

Small 5 mm left-sided intra canalicular vestibular schwannoma, which has not changed in
size since the last scan


##### #####
Consultant Radiologist
GMC ####
Ext: #####
```

Figure 2: *A radiology report sample of a positive tumour outcome.*

The reports were cleaned to remove text irrelevant to the modelling task, such as patient ID, date, clinician details, headers, and footers with administrative information. Statistical analysis of the cleaned

reports revealed that positive reports were lengthier on average than negative ones. In particular, the average number of words in the positive reports were 101.94 with standard deviation 39.55 compared to an average of 64.71 words in negative reports with standard deviation of 23.27. There was a total of 404,602 words in the cleaned corpus, with 19,131 unique words. Figure 3 shows a snippet of the word frequency, demonstrating that further cleaning of the data was required to remove stop words such as 'of', 'is', and 'the'. At the same time, an interesting finding was that the words 'auditory' and 'vestibular' had high frequency counts compared to those of stop words.

```
word:the, percentage:2.23
word:auditory, percentage:2.09
word:Both, percentage:2.01
word:vestibular, percentage:1.83
word:is, percentage:1.82
word:and, percentage:1.81
word:MRI, percentage:1.68
word:of, percentage:1.67
word:No, percentage:1.6
word:Findings:, percentage:1.49
```

Figure 3: *Word frequency result of the text corpus.*

In the task of detecting interesting or nuance findings and patterns in the corpus, EDA plays a crucial role in addition to unsupervised model. To gain further insights, the following strategies were employed:

- T-SNE was applied over the corpus to group the most prominent and similar words together to aid in capturing incidental findings or patterns and unusual/unexpected terms in the corpus. Word2Vec from genism.models was trained with the entire corpus to create token representations for each word with parameters size and min_count set to 300 and 100, respectively. T-SNE was executed with n_components = 2, init = pca, n_iter = 2500 and perplexity = 40.
- Five significant terms related to the context of brain tumours was selected for frequency analysis: 'schwannoma', 'vestibular', 'lesion', 'tumour' and 'findings'. The frequency of other words appearing with the five selected terms were calculated and a list of the top 20 was generated. For example, an expected high frequency pair would be 'vestibular schwannoma'. The word pairs were retrieved using NLTK ngram library and the frequency accumulated using 'collections.Counter'.
- The frequency between each pair of words was generated and the list of top 50 frequency pairs was filtered.
- The overall highest frequency words in the corpus were calculated and displayed graphically for further analysis.
- Whilst T-SNE could be used to get the synonyms of the five key terms identified by extracting words appearing near the search term, a more well-founded method was selected. In particular, gensim KeyedVectors provides a cosine-similarity function when working with word embeddings, which allows finding the cosine similarity of a term to all word vectors in the corpus. Moreover, BioWordVec capable of biomedical domain specific representations was used for word embedding over Word2Vec or Glove commonly used in T-SNE, thus providing better representations for clinical text. The results from the cosine similarities were sorted by descending order and the top 50 were filtered.

Data pre-processing

In NLP, data pre-processing involves stripping the raw text of all unwanted characters to ensure that no distortions are introduced into a subsequently trained model[18]. The following procedures were performed to pre-process the considered corpus:

- Simple text cleaning: regular expressions are used to remove special characters (e.g. ") and dealing with capitalization, punctuation signs, possessive pronouns, leading and trailing white spaces or tabs. In addition, domain specific terms that add noise to the model are identified and removed, these are 'findings', 'comment', 'erratum', 'mri internal auditory meatus both', 'clinical indication', 'conclusion', 'clinical history', 'impression' and so on;
- Stop-word removal: removing commonly used words such as 'the' that do not add much meaning to sentences;
- Lemmatization: transforming each word into its base form to correct non-real words in some cases; for example, transforming 'thu' into 'thus' through lemmatization.

LabelEncoder from the scikit-learn Python library was employed to encode positive and negative class

labels as 0 and 1, respectively, to ensure their compatibility with ML algorithms. The labels were used to validate the efficiency of label matching in unsupervised learning algorithms. The aim is to evaluate how well unsupervised learning algorithms could identify the positive and negative clusters.

Furthermore, given the unsupervised nature of the experiments, a deviation from the standard pre-processing techniques mentioned above were explored to understand whether negation by means of stop words such as 'no' and 'not' (i.e. maintaining the semantic meaning of sentences) would affect clustering performance. In line with the literature, this pre-processing stage had little to no impact on model performance[18]. Therefore, the standard pre-processing techniques were used.

Vector embeddings

Similar to class labels, many ML algorithms require categorical features to be converted into numeric form. There are many vectorization methods that could be employed in NLP for this step, each providing a different representation of the corpus: term frequency-inverse document frequency (TF-IDF), word embedding, contextual sub-word embedding and contextual string embedding. The aforementioned methods were explored in this study to determine which one achieves the best performance in unsupervised learning models. This step is important for accurate document representation because it impacts the calculation of the similarity measure between documents used to group them into clusters and identify meaningful patterns and diagnosis.

TF-IDF is known as the bag-of-words model, which ignores the word order in a document. It accounts for commonly occurring words in the document that generally do not contain useful or discriminatory information by down-weighing high frequency words[18]. It is important to note that TF-IDF implementations in Python libraries implicitly normalise the output using L2-normalisation[18]. Furthermore, TF-IDF vectors produce sparse matrices and require dimensionality reduction as a pre-processing step; principal component analysis (PCA) was used for this purpose. Preserving 95% of the variance, the TF-IDF matrices were reduced in dimension from shape (6117, 4898) to (6117, 1486).

Unlike TF-IDF, word embedding techniques produce low-dimensional floating-point vectors by capturing semantic relationships and syntactic similarity between words[5]. This study employed the most commonly used pre-trained word embedding algorithm, namely, Word2Vec[13]. As with any approach, this method is not without its limitations. For instance, the quality of the representations is impacted by out of vocabulary terms, due to its inability to capture the internal structure of words and domain specific representations. In a test transformation using Word2Vec, terms such as 'cerebellopontines', 'paranasal' and 'thecal' were not found in the vocabulary and were omitted. These limitations were addressed in the contextual sub-word embedding model, BioWordVec, which integrates domain knowledge in the training stage by merging sub-word data from biomedical text with the Medical Subject Headings (MeSH) vocabulary[25]. Another advancement from word embeddings, which provide distinct vectors per word, is a sentence embedding for biomedical text. Namely, BioSentVec demonstrated that it could capture sentence semantics, which the predecessors discussed could not[4]. These biomedical domain specific pre-trained embedding models are important to consider in clinical context as they could improve downstream ML tasks and are evaluated further in this study.

Deep learning models in NLP, such as BERT, ELMO, Roberta and OpenAI, to name a few, have advanced the field of NLP, producing state-of-the-art results comparable to those achieved in computer vision[7,11,16,17,22]. The ability to provide bi-directional representations 'jointly conditioned in both left and right contexts'[7] makes BERT uniquely suited to provide clinical report vector representations for unsupervised learning. This could improve the similarity scores between negative and positive tumour reports. Furthermore, as with BIOBERT and BioSentVec, biomedical and clinical variations of BERT were created to improve the performance of ML models in those domain specific contexts. More specifically, BIOBERT, BioClinicalBERT, BLUEBERT and BIOSYN are also compared in this study[1,9,15,21].

The final embedding considered in this study is a hybrid approach to generating contextual topic embeddings, known to considerably enhance performance by combining latent Dirichlet allocation (LDA) and contextual BERT embeddings[2,14,20]. However, since the combined vector is in a high dimensional space, autoencoders (AE) were used to learn lower dimensional latent space representational vectors. Hugging Face Sentence Transformers were used to achieve the embeddings, which provided an easy-to-use interface, reducing the effort required to generate the embedding of the clinical reports. The following embedding models were loaded using genism KeyedVectors:

- BioWordVec: pre-trained model (BioWordVec_PubMed_MIMICIII_d200.vec.bin) downloaded from https://github.com/ncbi-nlp/BioWordVec
- BioSentVec: pre-trained model (BioSentVec_PubMed_MIMICIII-bigram_d700.bin) downloaded from https://github.com/ncbi-nlp/BioSentVec

The following embedding models were loaded using Hugging Face Sentence Transformers:

- BERT: bert-base-uncased
- BioBERT: dmis-lab/biobert-v1.1
- BioBERT Lrg: dmis-lab/biobert-large-cased-v1.1
- BioSyn BioBert: dmis-lab/biosyn-biobert-bc2gn
- BioSyn SapBert: dmis-lab/biosyn-sapbert-bc5cdr-disease
- BlueBERT: bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12
- BioClinicalBERT: emilyalsentzer/Bio_ClinicalBERT

'BioBERT' and 'BioBERT Lrg', as referred to in this study, represent different versions of the pre-trained BIOBERT model, while 'BioSyn BioBert' and 'BioSyn SapBert' represent the biosyn variations of the models.

Model

Three well-known clustering algorithms were explored and compared to achieve the best performing vector embedding and unsupervised model combination using hard clustering. Namely, the centroid-based clustering algorithm, k-means[12], the balanced iterative reducing and clustering using hierarchies (BIRCH) algorithm[24] and the Gaussian mixture model (GMM)[19]. It is important to highlight that scikit-learn provides excellent defaults for their models, and as a result, the default settings were maintained throughout the experimentation process. The only parameter that was changed during experimentation was k, where it was set to either 2 or 3 clusters. Furthermore, the best performing model was measured by its ability to group clinical reports into positive and negative clusters; thereby facilitating an improved likelihood of identify patterns within the third cluster. Figure 4 depicts the modelling process, whereby each vector representation is combined with each of the clustering algorithms to produce the best performing model based on a set of metrics.

Metrics

Unsupervised learning methods is the core of this study, meaning that there were no annotated labels for the end result. However, the first stage of the experiments focuses on validation with the ground truth labels. The labels indicate whether a patient is positive for a tumour or not (negative result). These were used to identify how well the models perform in clustering the data into their respective groupings. Therefore, conventional classification metrics were used alongside clustering metrics. These include accuracy, precision, recall, f1-score and AUC. In addition, precision, recall and f1-score were calculated on both the positive and negative classes due to the high imbalanced nature of the data. Moreover, the ratio of negative to positive examples can have different effects on accuracy, and both precision and recall do not account for true negatives. The clustering metrics used in the experiments were Fowlkes–Mallows index, adjusted Rand index, silhouette coefficient and Davies–Bouldi index, as implemented in recent studies[6].

- Adjusted Rand index: a value close to 1 represents a perfect match whereas a value close to 0 represents random labelling.
- Fowlkes–Mallows index: ranges from 0 to 1. A high value indicates a good similarity between clusters.
- Silhouette coefficient: ranges from -1 to 1. Values close to the lower bound indicates wrong cluster assignment, close to 0 indicates overlapping clusters and 1 indicates perfect clustering.
- Davies–Bouldin index: The lower bound is 0, where lower values relate to a model with better separation between clusters.

Post clustering data augmentation and annotation correction

First correction: data augmentation

Given that the number of healthy people outweighs the number of those diagnosed with a specific condition, healthcare datasets are often not balanced in their class labels. If not addressed, the class imbalance problem can cause ML models to perform poorly. In unsupervised learning, this could mean the model is unable to cluster the majority class or both (the minority and majority class). This is because clustering algorithms are known to perform poorly when working with clusters of different densities[8]. This study con-

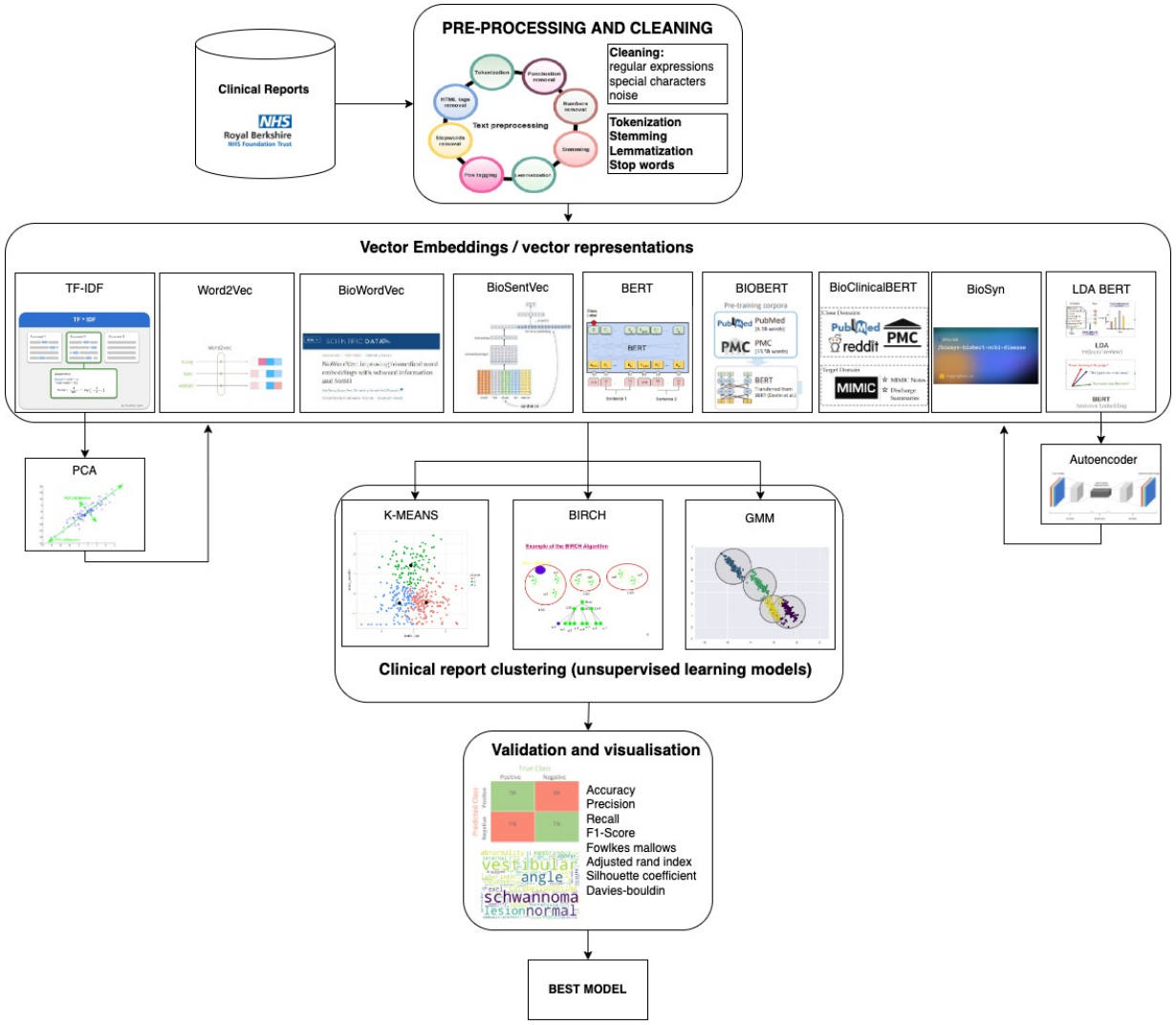

Figure 4: *General experimentation flow.*

sidered data augmentation to address the class imbalance problem and to understand its impact on the model's ability to create optimal clusters. The following three text augmentation strategies were explored in this study to increase the number of minority-class samples: The NLPAUG library was used to implement these data augmentation techniques.

- Synonym replacement: replacing some words with their synonyms[23]. The parameter aug_max was set to 3, indicating that a maximum of three words were replaced with their synonyms; and n was set to 2, indicating the generation of two augmented sentences from the original text.
- Word embedding replacement: replacing some words with their closest word vector in latent space using Glove, Word2Vec, EMLO or BERT[10].
- Back translation: creating slight variations by translating a document to another language and thereafter translating it back to the original language[3].

Word embedding replacement and backtranslation each create one augmented sentence. The process is depicted in Figure 5. The total number of positive samples after data augmentation was 4,700.

The general experimental flow shown in Figure 5 was then repeated with data augmentation included in the pre-processing step.

Second correction: annotation correction by expert

The addition of data augmentation to the pre-processing pipeline resulted in better clustering performance, where a greater number of samples were correctly clustered as negative and positive, as a result of balancing the respective cluster densities. Any further mismatch between labels were collected and shared with a clinician (domain expert) to validate whether the clustering algorithm was correct or the original annotations from the clinician. Moreover, the mislabeled samples (false negatives and positives) from all

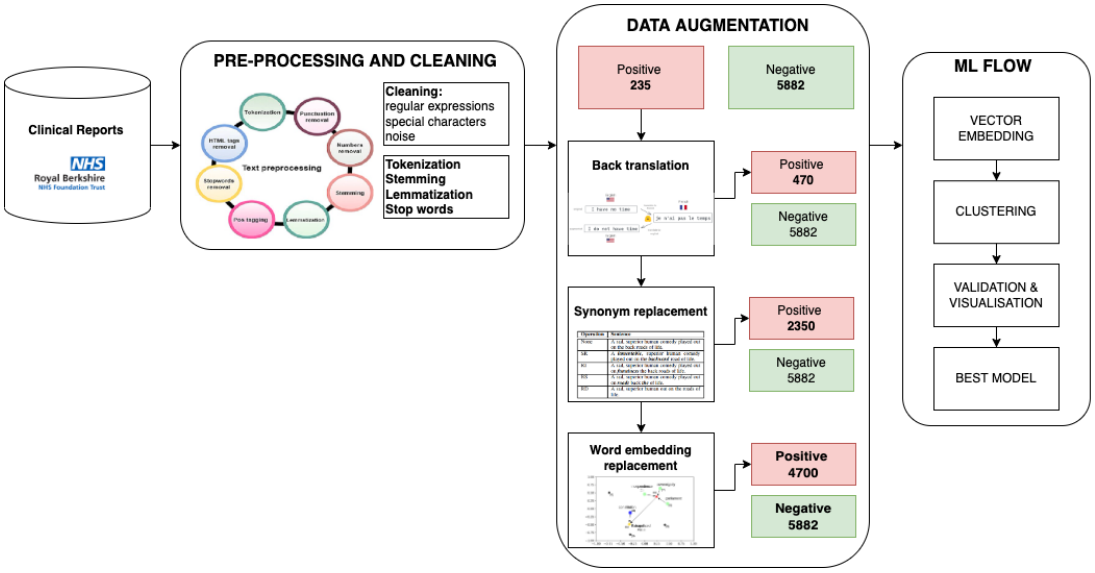

Figure 5: *Data augmentation process.*

the three clustering algorithms (k-means, BIRCH and GMM) were combined for this purpose. Figure 6 depicts the sequence of events occurring for the annotation correction by a clinical expert.

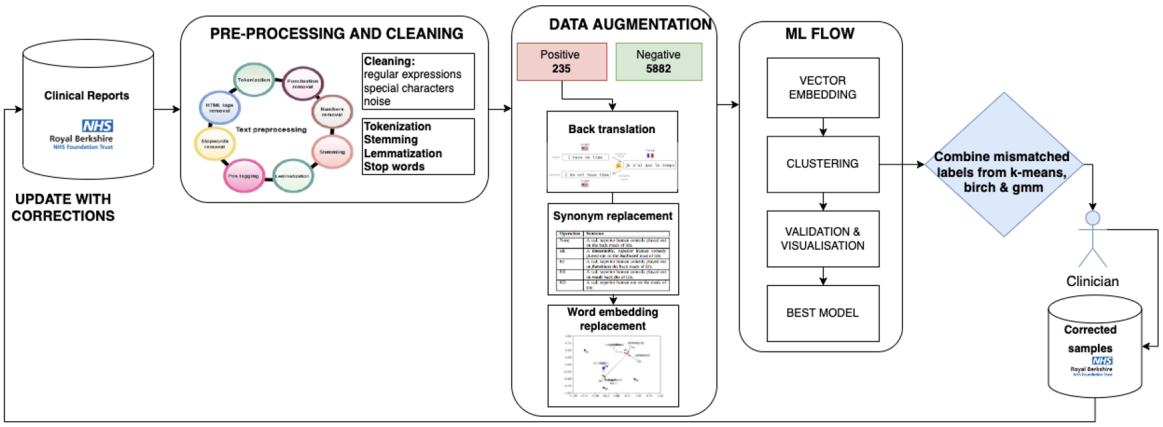

Figure 6: *Experimentation flow with second correction: annotation correction by expert.*

Summary of the experimentation process

Experiments 1 to 3 aimed to attain optimal clustering between the positive and negative groups when executed with 2 clusters. This was so that the best performing model had a greater chance of identifying patterns in a third cluster. Furthermore, the experiments were comparative in nature: finding the best vector representation that groups the data in their associated classes when clustered.

- **Experiment 1:** All vector embeddings were executed with each of the three unsupervised models and the results recorded for each combination (vector embedding + model).
- **Experiment 2:** The positive samples of the input data were augmented to increase the count of the minority class; thereafter experiment 1 was repeated to improve the clustering performance.
- **Experiment 3:** The mislabeled instances across all three best performing models and vector embedding combinations in experiment 2 were concatenated and corrected where applicable by a clinician. Data augmentation was performed again on the newly updated positive instances, and experiment 1 was repeated with the updated input.
- **Experiment 4:** The best performing vector embedding and model combinations from all three unsupervised models in experiment 3 were re-executed with 3 clusters. The aim was to find the ratio of positive and negative samples that contained additional information. This could be anything from interesting or vague reports to spotting patterns and incidental findings. The two out of three best performing model results of the third cluster were concatenated and shared with a clinician for further investigation.

Part II: Results

The results were progressive, in that one was building upon the other, improving the outcomes and likelihood of identifying the incidental findings.

Data pre-processing

Several data pre-processing techniques were applied to the original dataset in turn. The results of simple text cleaning using regular expressions and duplication removal are demonstrated in Figure 7. The total number of words decreased from 404,602 to 301,631, while the total number of unique words decreased from 19,131 to 5,516. It can be observed from Figure 7 that it is much easier to read the text after just one transformation. Figure 8 shows an example of text after removing stop words. The total number of words decreased by 85,608 words, and the total number of unique words decreased by 121 words. It can be noticed from Figure 9 that the high-frequency terms have become more relevant to the ML task after removing stop words. Furthermore, after applying lemmatization, the total number of words increased from 216,023 to 216,042 and the total number of unique words decreased from 5,395 to 4,923 (Figure 10). Moreover, whilst the above-mentioned pre-processing transformations have no impact on ML model performance, it is important to note that these transformations improve the computational complexity due to text reduction.

Original text

```
MR172016426 17/08/2017 MRI Internal auditory meatus Both\n\nClinical
History:\nAsymmetry in hearing, Left worse\\.br\\unilateral tinnitus, left
side, consistent tinnitus \\.br\\- Excl Vestibular Schwannoma\n\n\nStandard
sequences were performed, supplemented with an axial T2 sequence through
the brain.\n\nFindings:\nBoth internal auditory meati and membranous
labyrinths appear normal. \nNo cerebellopontine angle lesion or posterior
fossa abnormality is seen. In particular, no vestibular schwannoma is
identified.\nNo significant supratentorial abnormality is demonstrated.
\n\n\n\nDr ##### #####\nConsultant Radiologist\nGMC ######\n\n\n
```

Transformation

```
asymmetry in hearing left worse unilateral tinnitus left side consistent
tinnitus excl vestibular schwannoma standard sequences were performed
supplemented with an axial t2 sequence through the brain both internal
auditory meati and membranous labyrinths appear normal no cerebellopontine
angle lesion or posterior fossa abnormality is seen in particular no
vestibular schwannoma is identified no significant supratentorial
abnormality is demonstrated
```

Figure 7: *Example of text cleaning with regular expressions and duplication removal.*

Transformation

```
asymmetry hearing left worse unilateral tinnitus left side consistent
tinnitus excl vestibular schwannoma standard sequences performed
supplemented axial t2 sequence brain internal auditory meati membranous
labyrinths appear normal cerebellopontine angle lesion posterior fossa
abnormality seen particular vestibular schwannoma identified significant
supratentorial abnormality demonstrated
```

Figure 8: *Sample text after removing stop words.*

```
word:vestibular, percentage:3.93
word:schwannoma, percentage:3.82
word:normal, percentage:2.9
word:angle, percentage:2.69
word:lesion, percentage:1.9
word:hearing, percentage:1.67
word:abnormality, percentage:1.63
word:cerebellopontine, percentage:1.63
word:left, percentage:1.51
word:membranous, percentage:1.37
word:labyrinths, percentage:1.37
word:excl, percentage:1.36
word:posterior, percentage:1.34
word:fossa, percentage:1.32
word:internal, percentage:1.3
```

Figure 9: *Term frequency after removing stop words.*

Transformation

asymmetry hearing leave bad unilateral tinnitus leave side consistent tinnitus excl vestibular schwannoma standard sequence perform supplemented axial t2 sequence brain internal auditory meati membranous labyrinth appear normal cerebellopontine angle lesion posterior fossa abnormality see particular vestibular schwannoma identify significant supratentorial abnormality demonstrate

Figure 10: *Applying lemmatization: the transformed words are highlight in yellow.*

Exploratory data analysis

In the task of detecting interesting or nuance findings and patterns in the corpus, EDA plays a crucial role in addition to unsupervised models. To further elaborate, the frequency of words appearing with key terms were filtered and examined along with the synonyms. The selected key terms were 'schwannoma', 'vestibular', 'lesion, 'tumour', and 'findings'. A subset of these results for the terms 'schwannoma' and 'vestibular' are shown in Figure 11, while for the terms 'lesion' and 'tumour' are shown in Figure 12. In particular, an evaluation of the term 'schwannom' revealed that some terms occurring in the word pairs infer positive or negative outcomes, such as 'normal', 'exclude' or 'standard' are common in negative reports, whereas the terms 'right', 'internal' and 'intracanalicular' are frequent in positive reports. These can be further observed from the wordcloud results depicted in Figure 23. Furthermore, some identified synonyms appear as word pairs with 'schwannoma', such as 'cerebellopontin'. At the same time, the synonyms for the term 'vestibular' reveal some context from the study of the ear represented by terms such as 'retrocochlear', which refers to hearing loss, 'cochlea', which refers to a spiral shaped bone found in the ear, along with 'vertigo', and 'endolymphatic', among others. In addition, some terms related to the eye such 'nystagmus' and 'oculomotor' were also prevalent. Interestingly, since the term 'vestibular' often occurs before 'schwannoma' (with a frequency of 8,119), it's not surprising that some adjectives indicating a positive or negative outcome also appear as word pairs; such as 'excl', 'exclude', 'normal' indicate a negative outcome and 'demonstrate', 'leave', 'right', 'presence', 'symmetrical' possibly indicate a positive outcome.

pairs	frequency
('vestibular', 'schwannoma')	8119
('schwannoma', 'identify')	2118
('schwannoma', 'exclude')	1038
('schwannoma', 'internal')	557
('schwannoma', 'normal')	521
('schwannoma', 'cerebellopontine')	442
('schwannoma', 'acoustic')	383
('schwannoma', 'evidence')	218
('schwannoma', 'lt')	167
('schwannoma', 'technique')	159
('schwannoma', 'standard')	151
('schwannoma', 'rt')	115
('schwannoma', 'vestibular')	67
('schwannoma', 'prev')	67
('schwannoma', 'could')	51
('schwannoma', 'intracanalicular')	50
('schwannoma', 'central')	50
('schwannoma', 'pls')	47
('schwannoma', 'image')	46
('schwannoma', 'right')	41

words	schwannoma
paraganglioma	0.8097388
meningioma	0.80208325
neuroma	0.79584706
schawannoma	0.7942072
schannoma	0.78406394
lipoma	0.7676702
epidermoid	0.75160295
leiomyosarcoma	0.7432158
neurofibromatosis	0.73696864
cerebellopontine	0.7214702
schawnnoma	0.70857656
haemangioma	0.70831376
cavernoma	0.7066066
hemangioma	0.70219064
astrocytoma	0.7009963
neoplasm	0.69379765
chondromatosis	0.6920614
schwanomma	0.68259966
cylindroma	0.68077993
schwaanoma	0.6579058
cavernous	0.65678126
malignant	0.6566741
tumours	0.6558932
carcinomatosis	0.6533187
retrosigmoid	0.6508808
osteoma	0.6494224
exostosis	0.64112085
ecchordosis	0.6380137
intrasellar	0.63783246
pleomorphic	0.62703216
glomus	0.6261948
sellar	0.6224058
cholesteatomatous	0.6213327
adenoma	0.61862195
cyst	0.6166001
carcinoma	0.6164471
arachnoid	0.6161239

pairs	frequency
('vestibular', 'schwannoma')	8119
('excl', 'vestibular')	2476
('particular', 'vestibular')	2103
('demonstrate', 'vestibular')	1017
('exclude', 'vestibular')	566
('evidence', 'vestibular')	502
('bone', 'vestibular')	365
('neuroma', 'vestibular')	177
('leave', 'vestibular')	101
('right', 'vestibular')	79
('intracanalicular', 'vestibular')	76
('schwannoma', 'vestibular')	67
('feature', 'vestibular')	58
('vestibular', 'aqueduct')	47
('side', 'vestibular')	42
('sol', 'vestibular')	39
('presence', 'vestibular')	34
('sided', 'vestibular')	32
('normal', 'vestibular')	28
('symmetrical', 'vestibular')	28

words	vestibular
vestibulocochlear	0.8649392
vetibular	0.8286837
vestibulopathy	0.79860014
vestibualr	0.7886118
semicircular	0.76215637
vestibuar	0.75794274
retrocochlear	0.7509712
labyrinthine	0.74908763
oculomotor	0.74554163
auditory	0.7434534
neurosensory	0.7292815
intralabyrinthine	0.72666514
cochlear	0.7250955
labyrinths	0.7222029
trigeminal	0.70692927
vestib	0.7064318
brainstem	0.7016418
otolith	0.70053333
labyrinth	0.69577813
nystagmus	0.69173974
stapedial	0.68520725
cochlea	0.6849111
utricle	0.68368584
vertigo	0.6798169
vemps	0.6792392
vertiginous	0.6766748
sacculocollic	0.6761595
vemp	0.6756951
tinnitus	0.67479384
cerebellopontine	0.6724918
vestiblar	0.66760427
stapedius	0.66728806
hypoglossal	0.6669128
vhit	0.6651443
otological	0.6621878
endolymphatic	0.6605087
sensorineural	0.65345234

Figure 11: *Frequency of words occurring with schwannoma and synonyms of schwannoma (1st and 2nd columns) and frequency of words occurring with vestibular and synonyms of vestibular (3rd and 4th columns).*

Furthermore, T-SNE was executed on the corpus, and the results are shown in Figure 13. Feedback from

pairs	frequency
('angle', 'lesion')	3146
('lesion', 'posterior')	2247
('lesion', 'identify')	195
('mass', 'lesion')	172
('occupy', 'lesion')	160
('lesion', 'demonstrate')	137
('lesion', 'see')	129
('lesion', 'image')	75
('cpa', 'lesion')	62
('lesion', 'axial')	62
('lesion', 'detect')	46
('parenchymal', 'lesion')	45
('vi', 'lesion')	45
('t2', 'lesion')	43
('lesion', 'normal')	39
('enhance', 'lesion')	36
('lesion', 'leave')	33
('lesion', 'otherwise')	33
('lesion', 'right')	30
('lesion', 'measure')	29

words	lesion
nonenhancing	0.64840376
perilesional	0.63191456
nodule	0.6254211
lession	0.6217324
nodular	0.61124855
hyperintensity	0.59767675
lobulated	0.59577394
hyperintense	0.5952246
necrotic	0.58686036
circumscribed	0.586847
neoplasm	0.5815005
polypoid	0.5810366
infarcts	0.57985413
cyst	0.57442486
hemangioma	0.57392037
hyperenhancement	0.5725965
suspicious	0.5713333
cavernoma	0.56809086
hypointense	0.5662475
lipoma	0.5643766
pathologic	0.5609571
abnormality	0.55969024
stenosis	0.55458874
florid	0.55425876
scar	0.5539487
ectatic	0.55093974
focal	0.55041283
predilection	0.5500286
calcification	0.54943174
hyperintensities	0.5473334
stenosed	0.5455531
margin	0.5431755
incidental	0.54264617
excision	0.5417079
malignancy	0.5401497
radiologically	0.5400688
scalloping	0.53886956

pairs	frequency
('evidence', 'tumour')	14
('tumour', 'either')	12
('tumour', 'volume')	9
('enhance', 'tumour')	8
('tumour', 'measure')	7
('intracanalicular', 'tumour')	6
('cpa', 'tumour')	6
('tumour', 'appear')	5
('side', 'tumour')	5
('within', 'tumour')	5
('component', 'tumour')	5
('margin', 'tumour')	5
('sided', 'tumour')	4
('solid', 'tumour')	4
('base', 'tumour')	4
('increase', 'tumour')	4
('tumour', 'likely')	4
('angle', 'tumour')	4
('brain', 'tumour')	4
('tumour', 'normal')	4

words	tumour
metastasis	0.7079582
malignant	0.69438905
melanoma	0.6893492
metastatic	0.684995
carcinoma	0.6720268
neoplasm	0.6489572
astrocytoma	0.6376793
malignancy	0.63767165
adenoma	0.6364789
leiomyosarcoma	0.6322772
meningioma	0.625718
epidermoid	0.6223809
carcinomatosis	0.61571485
paraganglioma	0.61269945
necrosis	0.60292464
seminoma	0.5975407
schwannoma	0.59442246
schwanomas	0.58643705
schwannomas	0.58196324
cylindroma	0.58155936
cancer	0.5797054
macroadenoma	0.5708889
shwannoma	0.56316894
prolactinoma	0.561843
haemangioma	0.56152624
schwanoma	0.55720633
colorectal	0.5560634
squamous	0.547206
tissue	0.5427461
recurrence	0.5410587
lymphoma	0.53504866
perineural	0.5316266
radiotherapy	0.5290126
hemangioma	0.5266248
neurofibromatosis	0.5235126
schwannom	0.5159841
lipoma	0.5156654

Figure 12: *Frequency of words occurring with lesion and synonyms of lesion (1st and 2nd columns) and frequency of words occurring with tumour and synonyms of tumour (3rd and 4th columns).*

a clinician analysing these results confirmed the hypothesis that this kind of EDA may capture incidental findings: "I thought the results showed some promise in identifying scans with incidental findings - see the central cluster around 'cyst'". This can be seen in Figure 14. Moreover, a noteworthy finding is that a few terms in the central cluster of T-SNE identified by the clinician also appear in the synonyms for 'lesion'; for example, 'cyst', 'incidental' and 'focal'. In addition, the synonyms for the term 'tumour' contains words such as 'cancer' and 'necrosis', which are more specific for the positive finding and refer to the death of body tissue. It was noticed that the synonyms were very similar to those of the term 'vestibular', which appeared with a similarity score of 1.0. Other observations included that clinical radiologists often misspelled words and these misspellings filtered through the frequency results and so were removed. Furthermore, the top 50 most frequent word pairs (Figure 15) and the overall most frequent words (Figure 16) were retrieved, demonstrating an improvement in relevance after completing the pre-processing stage. These findings of identified synonyms and word pair frequencies would be valuable to clinicians in understanding the nature of the reports and patterns used by radiologists in recording the MRI findings. It can also be concluded that these findings provide meaningful information to some extent in the identification of incidental findings as demonstrated by the T-SNE results.

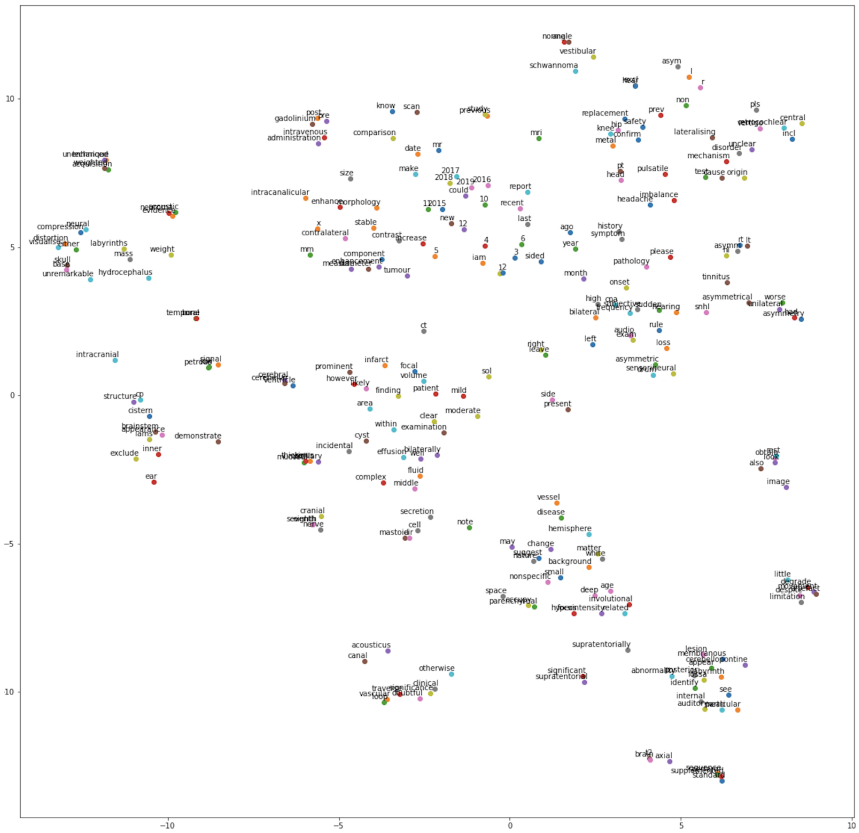

Figure 13: *T-SNE results of the clinical corpus (MRI reports), image file model_size_300_mincount_100.*

Figure 14: *T-SNE results zoomed in to the central region.*

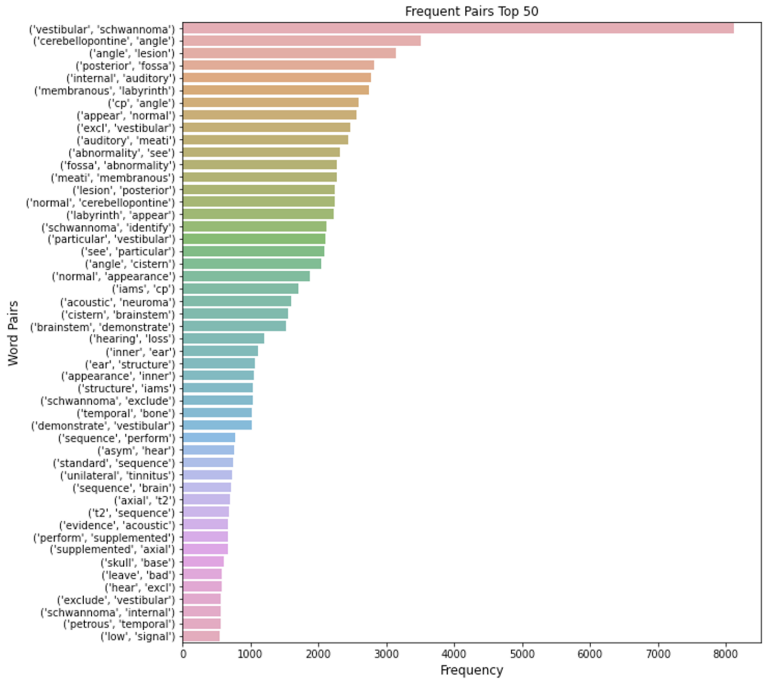

Figure 13 Frequent pairs top 50

Figure 15: *Frequent pairs top 50.*

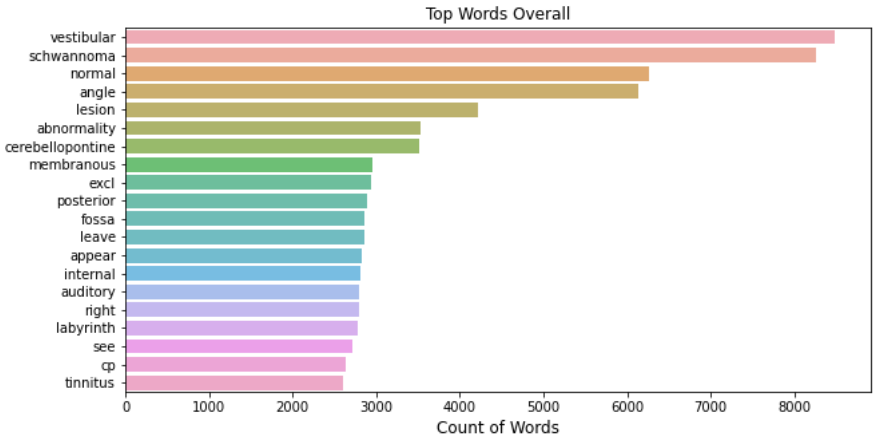

Figure 16: *Top overall words after pre-processing.*

Experiments with clustering models and embeddings

Experiments 1 to 3

The first three experiments aimed to identify the clustering framework that provided the most optimal clustering results, i.e. correctly grouping the positive and negative (for tumour) samples into their associated clusters. A large mismatch between instances and their ground truths were first addressed through data augmentation to improve the class imbalance and then clustered again. Any further mismatch between instances and their ground truths were then addressed by manual annotation of the mismatched instances to understand whether they were labelled incorrectly or clustered incorrectly. Moreover, these measures were taken to produce the best clusters for k=2, so that additional clustering would increase the likelihood of producing a third cluster with incidental findings, uncertainty, anomalies, or patterns in the cluster.

The first experiment executed k-means, Gaussian mixture model and BIRCH with k=2 clusters, for each vector embedding, and using the original dataset (6,117 samples - 5,882 negative and 235 positive). The results for k-means, GMM and BIRCH based clustering were sorted in descending order by their f1-scores and are shown in Tables Table 1, Table 2 and Table 3, respectively. K-means with BERT embedding produced the highest scores of 0.7893, 0.1454, 0.9191, 0.9959, 0.7841, 0.210 and 0.8774 for accuracy, precision, recall, and precision and recall on the negative class, respectively (herein referred to as precision neg and recall neg), f1-score and f1-score on the negative class (herein referred to as f1-score neg). This means that 1,270 out of 5,882 negative instances and 19 out of 235 positive instances were incorrectly clustered. Moreover, Bio Clinical BERT had the second highest scores of 0.1758, 0.7916, 0.6673, 0.0971, 0.9234 for f1-score, f1-score neg, accuracy, precision and recall, whereas word-embeddings, TF-IDF, BioSentVec and Word2Vec resulted in the lowest scores.

In contrast, BioSyn BioBert produced the highest scores for GMM clustering, with 0.1453, 0.7239, 0.5826, 0.0789, 0.9234, 0.9947 and 0.5690 for f1-score, f1-score neg, accuracy, precision, recall, precision neg and recall neg. This means that 2,535 out of 5,882 negative instances and 18 out of 235 positive instances were clustered incorrectly by GMM. Furthermore, BioBert resulted in the second highest scores and produced higher recall (0.9319) and precision neg (0.9950) scores compared to BioSyn BioBert. In addition, whilst BlueBERT and BioSyn SapBert resulted in higher accuracy scores of 0.8524 and 0.8576, respectively, they failed to cluster the positive class. BlueBERT reached 0.0103 precision and 0.0298 recall, misclustering 228 out of 235 positive instances and only 675 out of 5,882 negative instances; and BioSyn SapBert reached 0.0031 precision and 0.0085 recall, misclustering 233 out of 235 positive instances and only 638 out of 5,882 negative instances, which is the lowest misclustering for the negative class achieved thus far.

Concluding experiment 1, BIRCH resulted in the highest scores with Bio Clinical BERT (0.8545, 0.1780, 0.7702, 0.9894, 0.8579, 0.2891 and 0.9190 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively). This means that 836 out of 5,882 negative instances and 54 out of 235 positive instances were incorrectly clustered. Moreover, embeddings such as BERT, BioBERT, BlueBERT and other variations resulted in higher recall and precision neg scores compared to Bio Clinical BERT due to their ability to cluster the positive samples well. For example, BioBERT misclustered only 2 out of 235 positive instances but misclustered a high of 2,401 out of 5,882 negative instances. Interestingly, Word2Vec and BioSentVec consistently appeared in the lower performing range across all three clustering models. It is clear from the results that BIRCH with Bio Clinical BERT embeddings performed the best over k-means and GMM best performing models. However, the 890 out of 6,117 incorrectly clustered instances were attributed to the high imbalance between the positive and negative classes. It is well known that clustering algorithms perform poorly when some clusters have a higher density than others[8].

Experiment 2 showed that data augmentation improved the clustering results when dealing with highly imbalanced data. Experiment 1 was repeated with the augmented data (10,582 samples: 5,882 negative and 4,700 positive). The results are shown in Table 4, Table 5 and Table 6 for k-means, GMM and BIRCH with k=2 clusters, respectively. It is important to highlight that all augmented samples were omitted when calculating the evaluation metrics. All models reflected improved results across the board, with k-means producing its highest scores with LDA BERT embedding (0.9732, 0.5983, 0.9191, 0.9967, 0.9753, 0.7248 and 0.9859 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1 neg, respectively). In particular, only 145 out of 5,882 negative and 19 out of 235 positive samples were incorrectly clustered,

Table 1: Experiment 1: evaluation results of k-means for k=2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
BERT	0.7893	0.1454	0.9191	0.9959	0.7841	0.2510	0.8774	0.8009	0.1758	0.1412	2.4540
Bio Clinical BERT	0.6673	0.0971	0.9234	0.9954	0.6571	0.1758	0.7916	0.7225	0.1273	0.0533	2.6275
LDA BERT	0.6389	0.0932	0.9617	0.9976	0.6260	0.1699	0.7692	0.7095	0.5344	0.0422	0.7049
BioSyn SapBert	0.6194	0.0885	0.9574	0.9972	0.6059	0.1620	0.7538	0.7019	0.0878	0.0336	3.3391
BioWordVec	0.5463	0.0775	0.9915	0.9994	0.5286	0.1438	0.6914	0.6836	0.1474	0.0083	2.2120
BioSyn BioBert	0.5766	0.0778	0.9234	0.9946	0.5627	0.1435	0.7188	0.6892	0.1150	0.0169	2.7664
BlueBERT	0.5758	0.0768	0.9106	0.9937	0.5624	0.1416	0.7183	0.6890	0.1331	0.0164	2.4762
BioBert	0.5566	0.0751	0.9319	0.9950	0.5417	0.1390	0.7015	0.6852	0.1290	0.0110	2.4997
BioBert Lrg	0.5573	0.0702	0.8596	0.9898	0.5452	0.1298	0.7031	0.6853	0.1901	0.0101	1.8447
TF-IDF	0.5965	0.0013	0.0128	0.9402	0.6199	0.0024	0.7471	0.6959	0.0588	-0.0255	3.6658
BioSentVec	0.5271	0.0000	0.0000	0.9321	0.5481	0.0000	0.6903	0.6820	0.1266	-0.0119	2.5366
Word2Vec	0.6166	0.0000	0.0000	0.9414	0.6413	0.0000	0.7629	0.7027	0.1200	-0.0300	2.4695

Table 2: Experiment 1: evaluation results of GMM for k=2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
BioSyn BioBert	0.5826	0.0789	0.9234	0.9947	0.5690	0.1453	0.7239	0.6907	0.1152	0.0189	2.7756
BioBert	0.5534	0.0746	0.9319	0.9950	0.5383	0.1382	0.6986	0.6846	0.1285	0.0100	2.5024
BioBert Lrg	0.5578	0.0703	0.8596	0.9898	0.5457	0.1299	0.7036	0.6854	0.1900	0.0103	1.8449
BlueBERT	0.8524	0.0103	0.0298	0.9580	0.8852	0.0153	0.9202	0.8566	0.1686	-0.0368	2.0562
BioSyn SapBert	0.8576	0.0031	0.0085	0.9575	0.8915	0.0046	0.9233	0.8613	0.1151	-0.0457	2.8164
TF-IDF	0.5530	0.0012	0.0128	0.9358	0.5746	0.0022	0.7120	0.6854	0.0598	-0.0172	3.8090
BERT	0.4993	0.0011	0.0128	0.9293	0.5187	0.0020	0.6658	0.6806	0.0834	-0.0046	3.0802
Bio Clinical BERT	0.8511	0.0000	0.0000	0.9568	0.8851	0.0000	0.9195	0.8556	0.1230	-0.0500	2.2233
Word2Vec	0.6932	0.0000	0.0000	0.9475	0.7208	0.0000	0.8188	0.7387	0.0919	-0.0411	2.7032
LDA BERT	0.6882	0.0000	0.0000	0.9471	0.7157	0.0000	0.8153	0.7359	0.3758	-0.0405	0.7686
BioWordVec	0.6842	0.0000	0.0000	0.9468	0.7115	0.0000	0.8125	0.7337	0.0925	-0.0400	2.5846
BioSentVec	0.5271	0.0000	0.0000	0.9321	0.5481	0.0000	0.6903	0.6820	0.1266	-0.0119	2.5366

Table 3: Experiment 1: evaluation of BIRCH clustering for k=2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
Bio Clinical BERT	0.8545	0.1780	0.7702	0.9894	0.8579	0.2891	0.9190	0.8557	0.1300	0.1983	2.4141
BERT	0.7927	0.1465	0.9106	0.9955	0.7880	0.2524	0.8797	0.8036	0.1581	0.1436	2.7211
BioBERT	0.6072	0.0885	0.9915	0.9994	0.5918	0.1624	0.7434	0.6977	0.1103	0.0300	2.8223
BlueBERT	0.5936	0.0811	0.9277	0.9950	0.5802	0.1492	0.7330	0.6936	0.1140	0.0227	2.7838
BioSyn SapBert	0.5493	0.0780	0.9915	0.9994	0.5316	0.1446	0.6940	0.6840	0.0661	0.0092	3.9258
LDA BERT	0.5454	0.0771	0.9872	0.9990	0.5277	0.1430	0.6906	0.6834	0.4744	0.0080	0.7947
BioBert Lrg	0.5635	0.0768	0.9404	0.9957	0.5485	0.1420	0.7073	0.6864	0.1705	0.0131	1.9783
BioSyn BioBert	0.6132	0.0023	0.0213	0.9422	0.6369	0.0042	0.7600	0.7014	0.0869	-0.0276	2.7472
BioWordVec	0.8975	0.0025	0.0043	0.9591	0.9332	0.0032	0.9460	0.8984	0.2798	-0.0419	1.5034
TF-IDF	0.6081	0.0009	0.0085	0.9410	0.6321	0.0017	0.7562	0.6997	0.0528	-0.0278	4.2004
Word2Vec	0.9080	0.0000	0.0000	0.9594	0.9442	0.0000	0.9518	0.9085	0.2680	-0.0421	1.7405
BioSentVec	0.8779	0.0000	0.0000	0.9581	0.9130	0.0000	0.9350	0.8799	0.2981	-0.0479	1.7075

resulting in a reduction of 2,045 (+10 positive, -2,055 negative) misclustered samples compared to k-means with LDA-BERT in experiment 1, and a reduction of 1,125 (-1,125 negative) misclustered samples compared to the best performing k-means model in experiment 1 (BERT). Moreover, vector embeddings TF-IDF, Word2Vec and BioSentVec had a higher recall, showing improved clustering ability of the positive class whilst the quality of the negative cluster was drastically degraded. Interestingly, LDA BERT, Bio Clinical BERT and BioSyn SapBert consistently appeared across experiment 1 and 2 as higher performing embeddings for k-means.

AT the same time, GMM, one of the worst performing embeddings in experiment 1, achieved the best scores in experiment 2, i.e. BioWordVec with 0.9747, 0.6031, 0.9957, 0.9998, 0.9738, 0.7512 and 0.9867 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively. In particular, only 154 out of 5,882 negative and 1 out of 235 positive samples were incorrectly clustered, reducing the misclustered samples by 454 (-17 positive, -437 negative) samples. Whilst GMM provided a slightly improved clustering model with BioWord2Vec compared to k-means with LDA BERT, BIRCH with LDA BERT performed equally well achieving scores of 0.9748, 0.6057, 0.9872, 0.9995, 0.9743, 0.7508 and for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively. More specifically, BIRCH with LDA BERT only misclustered 151 out of 5,882 and 3 out of 235 positive samples, a huge improvement from experiment 1 with Bio Clinical BERT, reducing the misclustered samples by 736 (-51 positive, -685 negative) samples. Interestingly, Bio Clinical BERT was still among the top performing embeddings, resulting in the second highest scores for BIRCH based clustering in experiment 2. It can be observed from Table Table 6 that Word2Vec is not listed in the results; this is because removing the augmented samples resulted in a 0 count in the positive cluster (and 328 in the negative cluster), which disobeys the laws of clustering.

The results of experiments 1 and 2 are shown in Figure 17, Figure 18 and Figure 19. While data augmentation greatly improved the clustering performance, there was still 164, 155 and 154 samples incorrectly clustered after data augmentation for k-means, GMM and BIRCH, respectively. This means that either clustering or ground-truth labelling was incorrect. In addition, 139 samples were common across the misclustered samples for the best performing models for k-means, GMM and BIRCH. As a

result, the misclustered samples of each of the best performing models (k-means with LDA BERT, GMM with BioWordVec and BIRCH with LDA BERT) were concatenated, producing 185 potentially misclustered or mislabelled samples. These were sent to a clinician for further evaluation and verification.

The results received by the clinician confirmed that 143 out of 185 samples were mislabelled, meaning that 42 out of 185 samples were incorrectly clustered. In addition, 131 out of 139 mutually mislabelled (139 out of 185) samples were incorrectly labelled. Furthermore, manual annotation by an expert is a tedious and cumbersome task prone to errors. Thus, these results show that clustering could be used as an additional measure to verify manual annotations by an expert.

Table 4: Experiment 2: evaluation results of k-means clustering with k=2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
LDA BERT	0.9732	0.5983	0.9191	0.9967	0.9753	0.7248	0.9859	0.9714	0.5168	0.6904	0.8271
BioSyn SapBert	0.9279	0.3381	0.9149	0.9964	0.9284	0.4937	0.9612	0.9250	0.1606	0.4250	2.8135
Bio Clinical BERT	0.9295	0.3393	0.8809	0.9949	0.9315	0.4899	0.9622	0.9267	0.1946	0.4227	2.2217
BioWordVec	0.9194	0.3097	0.8936	0.9954	0.9204	0.4600	0.9565	0.9166	0.1703	0.3873	1.8552
Word2Vec	0.9062	0.2841	0.9489	0.9977	0.9045	0.4373	0.9488	0.9036	0.1070	0.3571	2.3680
BioSyn BioBert	0.8751	0.2225	0.9021	0.9955	0.8740	0.3569	0.9308	0.8742	0.1452	0.2666	2.6617
BERT	0.8900	0.2263	0.7702	0.9898	0.8948	0.3498	0.9399	0.8883	0.2161	0.2683	2.1031
BioSentVec	0.8436	0.1941	0.9745	0.9988	0.8383	0.3237	0.9115	0.8457	0.0744	0.2220	2.4511
BlueBERT	0.8444	0.1802	0.8596	0.9934	0.8438	0.2979	0.9125	0.8466	0.1223	0.2011	2.5996
BioBERT	0.8218	0.1663	0.9064	0.9955	0.8184	0.2810	0.8983	0.8272	0.1417	0.1775	2.5387
BioBert Lrg	0.6703	0.0905	0.8383	0.9904	0.6635	0.1634	0.7947	0.7241	0.1828	0.0484	1.8154
TF-IDF	0.4028	0.0600	0.9915	0.9991	0.3793	0.1131	0.5498	0.6962	0.0660	-0.0259	3.4125

Table 5: Experiment 2: evaluation results of GMM clustering with k=2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
BioWordVec	0.9747	0.6031	0.9957	0.9998	0.9738	0.7512	0.9867	0.9730	0.1983	0.7180	1.6549
Word2Vec	0.9631	0.5099	0.9872	0.9995	0.9621	0.6725	0.9804	0.9608	0.1261	0.6279	2.1714
TF-IDF	0.9480	0.4241	0.9872	0.9995	0.9464	0.5934	0.9722	0.9453	0.0560	0.5365	4.9708
Bio Clinical BERT	0.9465	0.4112	0.9064	0.9961	0.9481	0.5657	0.9715	0.9438	0.2082	0.5087	2.2095
BioSyn SapBert	0.9393	0.3811	0.9277	0.9969	0.9398	0.5403	0.9675	0.9365	0.1691	0.4782	2.7662
BioSentVec	0.9287	0.3493	0.9915	0.9996	0.9262	0.5166	0.9615	0.9258	0.0832	0.4471	2.4082
BlueBERT	0.9076	0.2794	0.8894	0.9952	0.9084	0.4252	0.9498	0.9051	0.1309	0.3471	2.7677
BioSyn BioBert	0.9004	0.2686	0.9234	0.9966	0.8995	0.4161	0.9456	0.8981	0.1532	0.3342	2.6961
BERT	0.8903	0.2289	0.7830	0.9904	0.8946	0.3542	0.9401	0.8886	0.2158	0.2724	2.1067
BioBERT	0.8669	0.2142	0.9234	0.9965	0.8647	0.3478	0.9259	0.8667	0.1466	0.2543	2.6429
LDA BERT	0.7082	0.1163	1.0000	1.0000	0.6965	0.2084	0.8211	0.7447	0.1806	0.0833	2.4211
BioBERT Lrg	0.6747	0.0921	0.8426	0.9907	0.6680	0.1660	0.7979	0.7263	0.1817	0.0508	1.8195

Experiment 3 updated the original dataset with the corrected samples by the clinician, resulting in 5,747 (reduced by 370) negative and 370 (increased by 135) positive samples. Moreover, experiments 1 and 2

Table 6: Experiment 2: evaluation of BIRCH clustering with k=2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
LDA BERT	0.9748	0.6057	0.9872	0.9995	0.9743	0.7508	0.9867	0.9731	0.5041	0.7178	0.8783
Bio Clinical BERT	0.9629	0.5088	0.9787	0.9991	0.9623	0.6696	0.9803	0.9606	0.2156	0.6249	2.2702
BioSyn BioBert	0.9599	0.4893	0.9702	0.9988	0.9595	0.6505	0.9788	0.9576	0.1654	0.6033	2.6817
BioSyn SapBert	0.9526	0.4474	0.9957	0.9998	0.9509	0.6174	0.9747	0.9500	0.1595	0.5640	2.9969
BioBERT	0.9441	0.4036	0.9532	0.9980	0.9437	0.5671	0.9701	0.9413	0.1583	0.5078	2.6223
BioSentVec	0.9263	0.3426	1.0000	1.0000	0.9233	0.5103	0.9601	0.9233	0.0754	0.4393	2.5033
BioWordVec	0.9232	0.3300	0.9702	0.9987	0.9213	0.4924	0.9584	0.9203	0.1635	0.4202	2.0281
BERT	0.8934	0.2644	0.9957	0.9998	0.8893	0.4179	0.9413	0.8913	0.1888	0.3313	2.3638
BlueBERT	0.6469	0.0978	0.9957	0.9997	0.6329	0.1781	0.7751	0.7128	0.1097	0.0479	2.8505
BioBERT Lrg	0.5197	0.0741	1.0000	1.0000	0.5005	0.1379	0.6671	0.6810	0.1649	0.0003	2.0252
TF-IDF	0.4085	0.0603	0.9872	0.9987	0.3854	0.1137	0.5562	0.6944	0.0635	-0.0246	3.5160

were repeated in experiment 3 to evaluate any improvement in the clustering performance. The results (without data augmentation) are depicted in Table 7, Table 8 and Table 9 for k-means, GMM and BIRCH, respectively. Interestingly, the embeddings that performed well in experiment 1 are still among the top four in experiment 3. For example, similarly, to experiment 1, BERT, Bio Clinical BERT, LDA BERT and BioSyn SapBert are the best performing embeddings for k-means; this is also true for BIRCH. Furthermore, the results demonstrated that 1,158, 1,903 and 781 instances were misclustered compared to 1,289, 609 and 890 achieved in experiment 1 for k-means, GMM and BIRCH, respectively. Whilst the results show some improvement in k-means and BIRCH, the results for GMM are drastically worse. It is worthwhile noting that k-means and BIRCH resulted in the same vector embedding for the best performing model (BERT and Bio Clinicial BERT) in experiment 1 and 3, whereas GMM differed from the results in experiment 1 (BioSyn BioBert), producing Bio Clinical BERT in experiment 3. These results dis not help achieve a high match between samples and their ground truths; there was still a high number of misclustered instances. Therefore, data augmentation was used to improve the clustering results again. The data had to be re-augmented because the original dataset was updated with the correct labels, increasing the count on the positive class (from 235 to 370). After data augmentation there are 10,187 samples, of which 5,747 are negative and are 4,440 positive.

Data augmentation proved to increase the clustering performance again, resulting in the same vector embeddings as those in experiment 2. However, the results were measurably improved. K-means with LDA BERT achieved 0.9946, 0.9720, 0.9378, 0.9960, 0.9983, 0.9546 and 0.9971 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively. In particular, there were only 10 out of 5,747 negative and 23 out of 370 positive misclustered instances. In addition, GMM with BioWordVec achieved scores of 0.9928, 0.9015, 0.9892, 0.9993, 0.9930, 0.9433 and 0.9962 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively. This means that 40 out of 5,747 negative and 4 out of 370 samples that were incorrectly clustered. Moreover, BIRCH with LDA BERT achieved the highest scores (0.9953, 0.9273, 1, 1, 0.9950, 0.9623 and 0.9975 for accuracy, precision, recall, precision neg, recall neg, f1-score and f1-score neg, respectively), with only 29 out of 5,747 negative and 0 out of 370 positive misclustered instances. In addition, BIRCH achieved the highest Fowlkes-Mallows score (0.9947), silhouette coefficient (0.4902), adjusted Rand score (0.9546) and the lowest Davies–Bouldin score (0.9230); these are used to determine the goodness measure of the clusters, evaluating the separation within and between the clusters.

The results with and without data augmentation can be visually interpreted as shown in Figure 20, Fig-

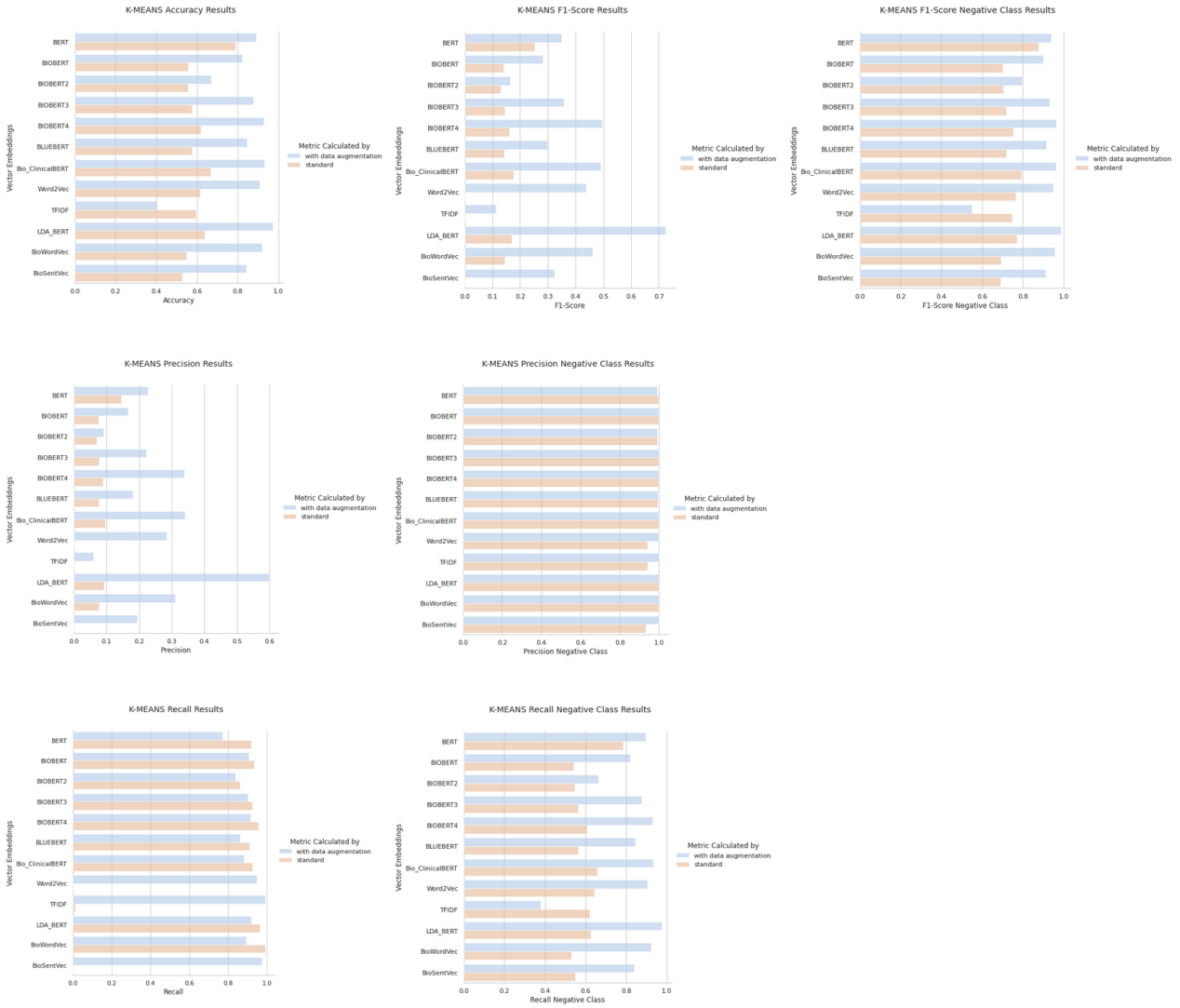

Figure 17: *Evaluation metrics of k-means clustering with and without data augmentation for all vector representations.*

ure 21 and Figure 22. The experiments showed that the results after ground-truth correction and data augmentation improved the model's ability to group the data into positive and negative clusters. As a result, the best performing models of experiment 3 were used in experiment 4, since these experiments demonstrated that clustering could be used instead of classification. In addition, a word cloud was generated from the best performing model in this experiment (BIRCH with LDA BERT embedding), which depicts the high frequency terms in each cluster (Figure 23). Moreover, feedback on these generated word clouds were received from a clinician who advised the following:

- TOPIC0_WORDCLOUD (negative cluster)
  "Vestibular schwannomas occur in the cerebellopontine angle – good. There isn't a side preponderance in affected patients; I'd expect right and left to be equally weighted. I cannot find 'left' but I can find 'leave'. This is an error that needs be corrected."
- TOPIC1_WORDCLOUD (positive cluster)
  "Same again with 'leave' instead of 'left'. Otherwise, the weightings look appropriate. I did think 'X'? But of course, this is from dimensions - 10mm x 15mm."

The issue with the word 'leave' instead of 'left' originated from lemmatisation in the pre-processing step and could be avoided early on. In addition, it is interesting that dimensions representing size of a tumour correctly appear dominant in the positive cluster.

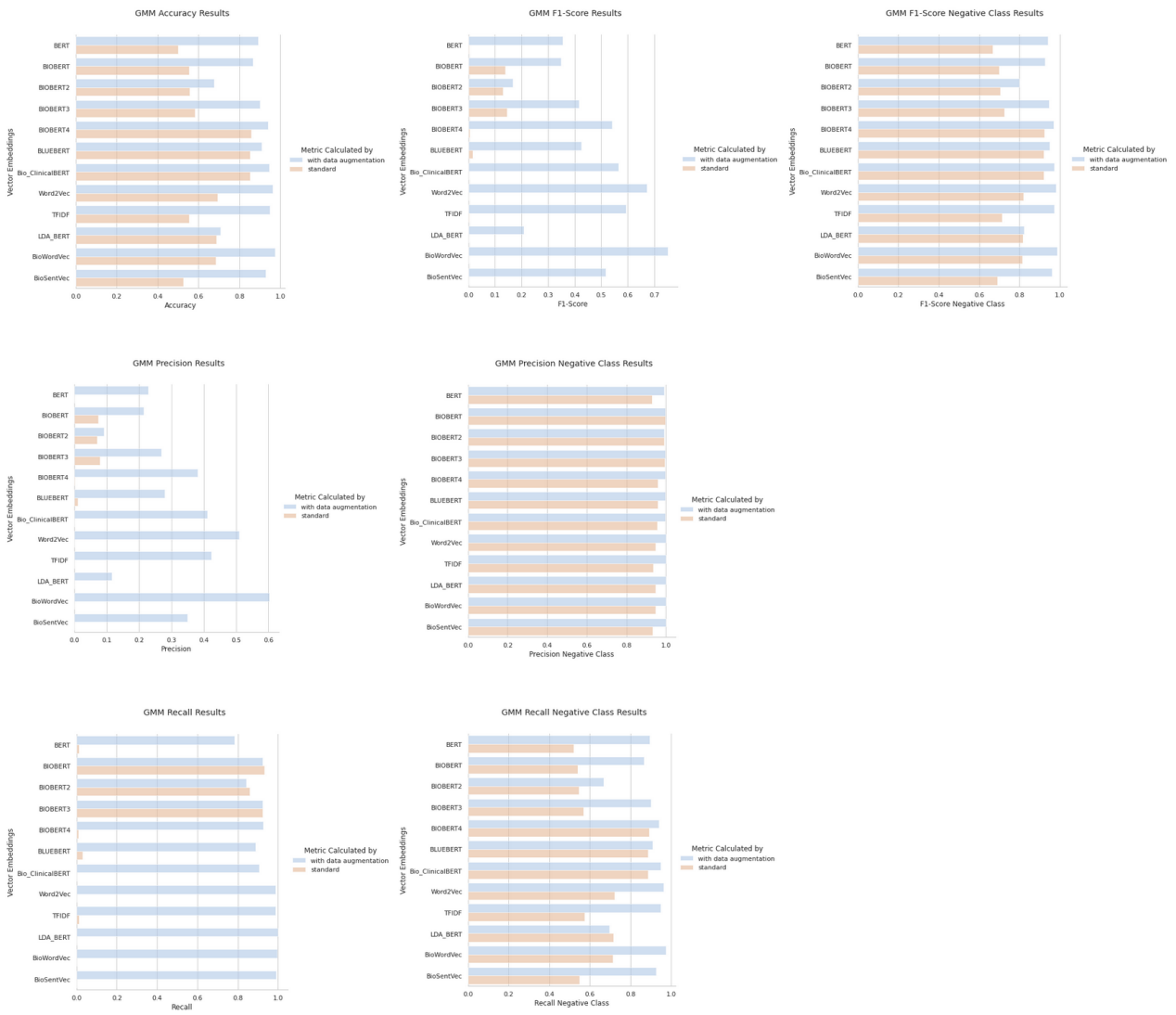

Figure 18: *Evaluation metrics of Gaussian mixture model clustering with and without data augmentation for all vector representations.*

Experiment 4

Experiment 4 aimed to verify whether introducing a third cluster given two known classes would provide further insight into reports by spotting anomalies or incidental findings. In other words, determine whether the top performing models in experiment 3 that successfully clustered the data into two classes could find a third cluster of patterns other than positive or negative. The third cluster should end up with positive and negative samples that contain other patterns in the data; this could be incidental findings, uncertainty in the reports or any other interesting findings. Each of the best performing models from experiment 3 was re-executed with k=3 clusters. Furthermore, the evaluation metrics used for this experiment were the silhouette coefficient and Davies–Bouldin index. There are no ground truth labels for the third cluster, therefore the other evaluation metrics could not be used. Furthermore, the models with data augmentation from experiment 3 were used with k-3 clusters; however, the metrics were measured without the augmented samples. Table 13 shows the resulting number of samples in each cluster, while Table 14 shows the number of samples that match with the ground truth for each cluster for k-means with LDA BERT and k=3 clusters. Only 8 samples in the negative and 1 in the positive cluster were incorrectly clustered. Furthermore, 109 positive and 24 negative samples were identified in the third cluster.

In contrast, GMM did not perform as well. The results are listed in Table 15 and Table 16, with 85 negative and 2 positive samples incorrectly clustered. Furthermore, 1 positive and 1,686 negative samples were identified in the third cluster. Interestingly, BIRCH with LDA BERT and k=3 clusters performed similarly to k-means, as shown in Table 17 and Table 18, where 0 samples in the negative and only 1 sample in

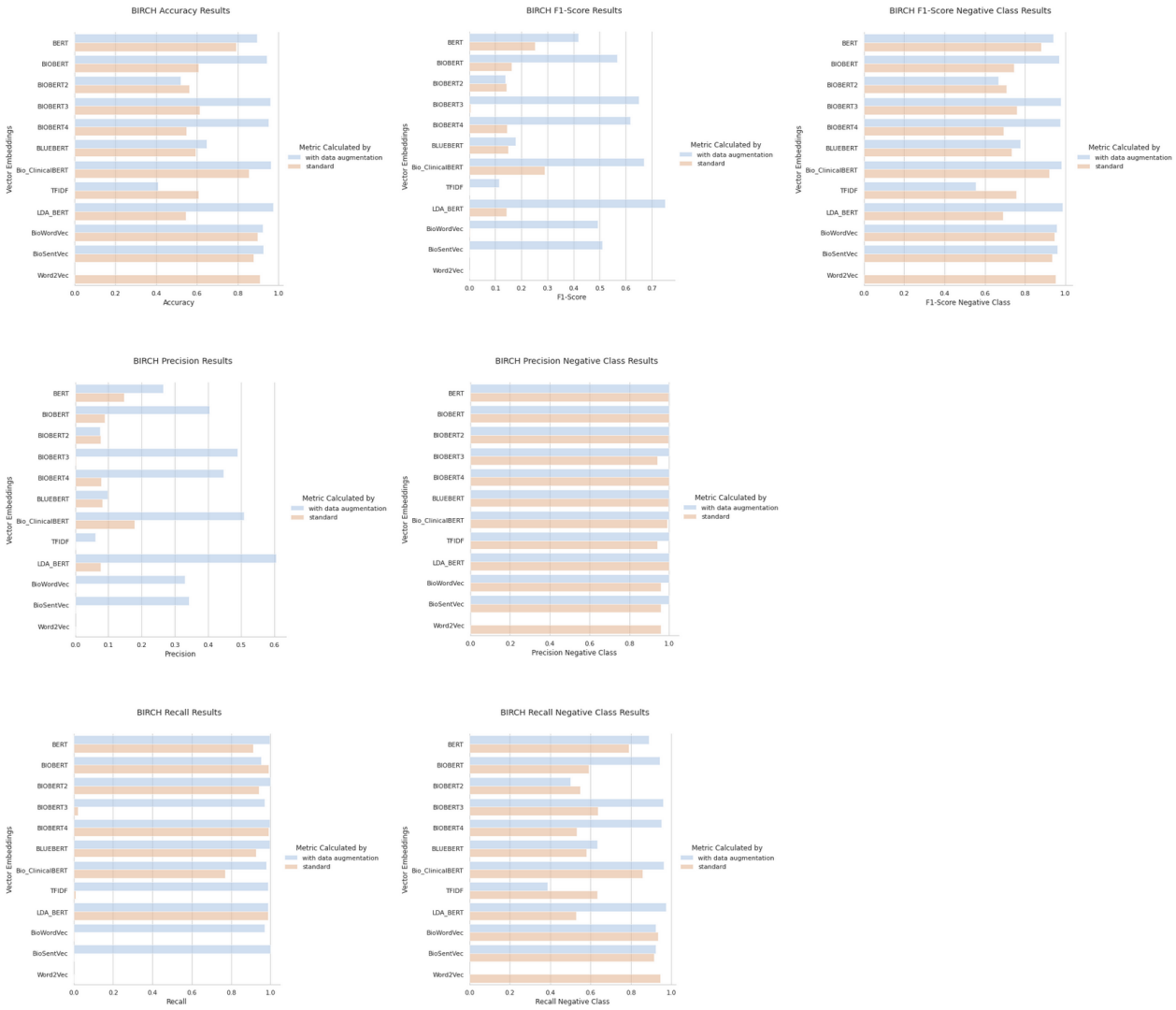

Figure 19: *Evaluation metrics of BIRCH clustering with and without data augmentation for all vector representations.*

the positive cluster were incorrectly clustered. In addition, 108 positive and 28 negative samples were identified in the third cluster. K-means and BIRCH resulted in very similar outcomes, whereas GMM with Word2Vec achieved completely different results. For example, k-means and BIRCH resulted in less than 137 samples in the third cluster, whereas GMM contained 1,687 samples in the third cluster. Moreover, there were more positive samples in the third cluster for k-means and BIRCH (109 out of 133 and 108 out of 136) compared to GMM, with a majority of 1,686 out of 1,687 negative samples.

Table 19 shows the clustering performance of the models. These results indicate the goodness measure of the clusters. A silhouette score close to 1 indicates well defined clusters, and a lower DB score indicates better separation between clusters. K-means with LDA BERT resulted in the highest silhouette coefficient (0.3616), while BIRCH with LDA BERT desirably resulted in the lowest Davies–Bouldin score (1.3887). In contrast, GMM with BioWordVec produced the worst silhouette coefficient (0.0841), where a value close to 0 is indicative of overlapping clusters, and the highest Davies–Bouldin score of 2.366. These results indicate that k-means and BIRCH resulted in the best performing models for k=3 clusters, therefore the GMM model was eliminated for further processing.

It is important to highlight that these models performed better than the models produced by (Davagdorg, Park, et.al, 2022) (0.3041 and 1.85 for silhouette coefficient and Davies–Bouldin index, respectively), whose research formed the basis for the existing methodology.

The outputs of the third clusters from k-means and BIRCH were concatenated, resulting in a total of 158 samples, of which 115 were mutual samples shared across both model results. In addition, the BIRCH with LDA model had the least mismatch between the two classes, even with 3 clusters, resulting in the

Table 7: Experiment 3: evaluation of results with corrected samples k-means clustering with 2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
BERT	0.8107	0.2349	0.9432	0.9955	0.8022	0.3761	0.8884	0.8092	0.1758	0.2285	2.4540
Bio Clinical BERT	0.6871	0.1544	0.9324	0.9936	0.6713	0.2650	0.8012	0.7198	0.1273	0.0888	2.6275
LDA BERT	0.6609	0.1488	0.9757	0.9976	0.6407	0.2582	0.7803	0.7057	0.5344	0.0729	0.7049
BioSyn SapBert	0.6394	0.1406	0.9703	0.9969	0.6181	0.2456	0.7631	0.6956	0.0876	0.0578	3.3407
BioSyn BioBert	0.5977	0.1251	0.9432	0.9937	0.5754	0.2210	0.7288	0.6803	0.1150	0.0323	2.7664
BioWordVec	0.5678	0.1218	0.9892	0.9987	0.5406	0.2168	0.7015	0.6725	0.1474	0.0181	2.2120
BlueBERT	0.5952	0.1223	0.9216	0.9913	0.5742	0.2160	0.7272	0.6795	0.1331	0.0304	2.4762
BioBERT	0.5772	0.1198	0.9432	0.9934	0.5537	0.2125	0.7111	0.6747	0.1290	0.0221	2.4997
BioBert Lrg	0.5686	0.1057	0.8216	0.9796	0.5523	0.1872	0.7064	0.6728	0.1901	0.0161	1.8447
TF-IDF	0.5745	0.0013	0.0081	0.9054	0.6109	0.0023	0.7296	0.6770	0.0588	-0.0353	3.6658
Word2Vec	0.5949	0.0005	0.0027	0.9079	0.6330	0.0008	0.7460	0.6832	0.1200	-0.0414	2.4695
BioSentVec	0.5053	0.0004	0.0027	0.8933	0.5377	0.0007	0.6713	0.6664	0.1266	-0.0134	2.5366

best performing model for this experiment.

The wordcloud for the three clusters are shown in Figure 24. Interestingly, the word frequency of positive and negative samples identified in the respective wordclouds are proportionate to the positive and negative samples found in the third cluster.

It can be concluded from this experiment that introducing a third cluster is potentially useful for spotting anomalies in clinical reports (for example, due to errors made by clinicians) or registering the progression of patients. For example, this could be cases, where a tumour could not be captured in previous MRI scans (Figure 25) but become obvious in follow-ups, or vice versa, a tumour was cured, and reports mention both the history of the patient and their current state (Figure 26). Figure 25 shows the former observation, where the Clinical History section reported "exclude vestibular schwannoma"; however, the Findings section provide a detailed description of a lesion measuring up to 1.3cm.

The clustering results were visualised using UMAP (Figure 27). It is clear from the figure that the hybrid, LDA and BERT, approaches provide a better separation between clusters for both k-means and BIRCH, demonstrating a high inter- and low intra-cluster distance, whereas the converse is true for GMM, which demonstrated overlapping clusters and a spread-out distribution of the clusters.

References

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings. 4 2019.

[2] Pranjali Basmatkar and Mahesh Maurya. An overview of contextual topic modeling using bidirectional encoder representations from transformers. *Lecture Notes in Electrical Engineering*, 844:489–504, 2022.

[3] Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. Back translation,cyberbullying detection,encoder–decoder,hate speech,nlp transformers,paraphrasing. *Online Social Networks and Media*, 24:100153, 2021.

[4] Qingyu Chen, Yifan Peng, and Zhiyong Lu. Biosentvec: Creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, 6 2019.

[5] F Chollet. *Deep Learning with Python*. Manning Publications, New York, 2018.

[6] Khishigsuren Davagdorj, Ling Wang, Meijing Li, Van-Huy Pham, Keun Ho Ryu, and Nipon Theera-

Table 8: Experiment 3: evaluation of results with corrected samples Gaussian mixture model clustering with 2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
Bio Clinical BERT	0.6889	0.1552	0.9324	0.9936	0.6732	0.2661	0.8026	0.7209	0.1275	0.0902	2.6273
BioSyn BioBert	0.6047	0.1274	0.9459	0.9941	0.5827	0.2245	0.7348	0.6825	0.1153	0.0361	2.7754
BioBERT	0.5738	0.1189	0.9432	0.9934	0.5500	0.2112	0.7080	0.6739	0.1285	0.0205	2.5024
BioBert Lrg	0.5691	0.1058	0.8216	0.9796	0.5528	0.1874	0.7068	0.6729	0.1900	0.0163	1.8449
BlueBERT	0.8342	0.0292	0.0541	0.9356	0.8845	0.0380	0.9093	0.8372	0.1686	-0.0354	2.0600
BioSyn SapBert	0.8359	0.0047	0.0081	0.9330	0.8892	0.0059	0.9105	0.8393	0.1151	-0.0618	2.8164
BERT	0.4775	0.0014	0.0108	0.8885	0.5076	0.0025	0.6461	0.6666	0.0834	-0.0022	3.0802
TF-IDF	0.5321	0.0012	0.0081	0.8986	0.5659	0.0021	0.6944	0.6687	0.0597	-0.0224	3.8073
Word2Vec	0.6714	0.0006	0.0027	0.9175	0.7145	0.0010	0.8034	0.7172	0.0919	-0.0578	2.7032
BioWordVec	0.6624	0.0006	0.0027	0.9165	0.7049	0.0010	0.7969	0.7123	0.0925	-0.0562	2.5846
BioSentVec	0.5777	0.0005	0.0027	0.9054	0.6148	0.0008	0.7323	0.6780	0.1241	-0.0368	2.6161
LDA BERT	0.4958	0.0000	0.0000	0.8913	0.5278	0.0000	0.6630	0.6662	0.3650	-0.0099	0.9861

Umpon. Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering. *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 5893*, 19:5893, 5 2022.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Naacl-Hlt 2019*, 1:4171–4186, 2018.

[8] Aurelienon Ger. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Incorporated, 2019.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 1 2019.

[10] Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. A survey of text data augmentation. *Proceedings - 2020 International Conference on Computer Communication and Network Security, CCNS 2020*, pages 191–195, 2020.

[11] 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.

[12] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1 2013.

[14] Sarojadevi Palani, Prabhu Rajagopal, and Sidharth Pancholi. T-bert – model for sentiment analysis of micro-blogs integrating topic model and bert. 6 2021.

[15] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language pro-

Table 9: Experiment 3: evaluation of results with corrected samples BIRCH clustering with 2 clusters

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
Bio Clinical BERT	0.8723	0.2979	0.8189	0.9869	0.8758	0.4369	0.9280	0.8660	0.1300	0.3206	2.4141
BERT	0.8131	0.2355	0.9297	0.9944	0.8056	0.3758	0.8901	0.8114	0.1581	0.2299	2.7211
BioBERT	0.6286	0.1390	0.9892	0.9989	0.6054	0.2437	0.7538	0.6911	0.1103	0.0520	2.8223
BlueBERT	0.6157	0.1314	0.9541	0.9950	0.5939	0.2309	0.7438	0.6862	0.1140	0.0426	2.7838
BioSyn SapBert	0.5714	0.1232	0.9946	0.9994	0.5441	0.2192	0.7046	0.6733	0.0661	0.0199	3.9258
BioBert Lrg	0.5853	0.1234	0.9595	0.9954	0.5612	0.2187	0.7177	0.6767	0.1705	0.0264	1.9783
LDA BERT	0.5674	0.1219	0.9919	0.9990	0.5401	0.2172	0.7012	0.6725	0.4744	0.0180	0.7947
BioWordVec	0.8758	0.0051	0.0054	0.9357	0.9318	0.0052	0.9337	0.8767	0.2798	-0.0524	1.5034
BioSyn BioBert	0.5915	0.0028	0.0162	0.9085	0.6285	0.0048	0.7430	0.6820	0.0869	-0.0388	2.7472
BioSentVec	0.8561	0.0020	0.0027	0.9342	0.9111	0.0023	0.9225	0.8581	0.2981	-0.0610	1.7075
TF-IDF	0.5890	0.0009	0.0054	0.9073	0.6266	0.0016	0.7413	0.6813	0.0540	-0.0395	4.1269
Word2Vec	0.8859	0.0000	0.0000	0.9361	0.9429	0.0000	0.9395	0.8867	0.2680	-0.0526	1.7405

Table 10: Experiment 3: evaluation of results with corrected samples k-means with 2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
LDA BERT	0.9946	0.9720	0.9378	0.9960	0.9983	0.9546	0.9971	0.9940	0.5127	0.9460	0.8346
BioSyn SapBert	0.9567	0.5904	0.9270	0.9951	0.9586	0.7213	0.9765	0.9522	0.1660	0.6637	2.7430
Bio Clinical BERT	0.9549	0.5822	0.9000	0.9933	0.9584	0.7070	0.9756	0.9503	0.2002	0.6478	2.2203
BlueBERT	0.9496	0.5566	0.8243	0.9883	0.9577	0.6645	0.9728	0.9449	0.1525	0.6010	2.5551
BioWordVec	0.9400	0.5022	0.9054	0.9936	0.9422	0.6461	0.9672	0.9344	0.1724	0.5724	1.8466
Word2Vec	0.9273	0.4525	0.9649	0.9976	0.9248	0.6160	0.9598	0.9209	0.1085	0.5304	2.3740
BioSyn BioBert	0.9029	0.3758	0.9162	0.9941	0.9020	0.5330	0.9458	0.8959	0.1484	0.4305	2.6591
BERT	0.9156	0.4000	0.7892	0.9855	0.9238	0.5309	0.9537	0.9093	0.2207	0.4404	2.0620
BioSentVec	0.8697	0.3142	0.9757	0.9982	0.8629	0.4753	0.9256	0.8630	0.0740	0.3514	2.4437
BioBERT	0.8424	0.2691	0.9351	0.9950	0.8364	0.4179	0.9089	0.8374	0.1417	0.2828	2.5627
BioBert Lrg	0.6538	0.1309	0.8378	0.9840	0.6419	0.2264	0.7770	0.7026	0.1860	0.0574	1.8274
TF-IDF	0.4259	0.0949	0.9946	0.9991	0.3892	0.1733	0.5602	0.6770	0.0658	-0.0355	3.4223

cessing: An evaluation of bert and elmo on ten benchmarking datasets. *BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 6 2019.

[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and

Table 11: Experiment 3: evaluation of results with corrected samples Gaussian mixture model with 2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
BioWordVec	0.9928	0.9015	0.9892	0.9993	0.9930	0.9433	0.9962	0.9919	0.1987	0.9318	1.6836
Word2Vec	0.9702	0.6709	0.9973	0.9998	0.9685	0.8022	0.9839	0.9669	0.1165	0.7595	2.3507
TF-IDF	0.9694	0.6661	0.9919	0.9995	0.9680	0.7970	0.9835	0.9660	0.0577	0.7534	4.9817
Bio Clinical BERT	0.9675	0.6654	0.9297	0.9954	0.9699	0.7756	0.9825	0.9639	0.2086	0.7301	2.2178
BioSyn SapBert	0.9627	0.6282	0.9405	0.9960	0.9642	0.7532	0.9798	0.9587	0.1702	0.7021	2.7236
BlueBERT	0.9621	0.6420	0.8432	0.9897	0.9697	0.7290	0.9796	0.9582	0.1604	0.6781	2.5298
BioSentVec	0.9392	0.4986	0.9865	0.9991	0.9361	0.6624	0.9666	0.9334	0.0806	0.5868	2.4333
BioSyn BioBert	0.9219	0.4322	0.9297	0.9951	0.9214	0.5901	0.9568	0.9153	0.1545	0.5007	2.6901
BERT	0.9160	0.4030	0.8081	0.9868	0.9229	0.5378	0.9538	0.9096	0.2203	0.4470	2.0734
BioBERT	0.8867	0.3427	0.9514	0.9965	0.8825	0.5039	0.9361	0.8796	0.1464	0.3903	2.6707
LDA$_BERT$	0.7160	0.1756	1.0000	1.0000	0.6978	0.2987	0.8220	0.7373	0.1530	0.1209	2.6709
BioBert Lrg	0.6598	0.1349	0.8541	0.9857	0.6473	0.2330	0.7814	0.7055	0.1847	0.0625	1.8343

Table 12: Experiment 3: evaluation of results with corrected samples BIRCH with 2 clusters and data augmentation

Model	Acc	Prec	Recall	Prec Neg	Recall Neg	F1	F1 Neg	FMS	SC	ARS	DBS
LDA BERT	0.9953	0.9273	1.0000	1.0000	0.9950	0.9623	0.9975	0.9947	0.4902	0.9546	0.9230
TF-IDF	0.9949	0.9775	0.9378	0.9960	0.9986	0.9572	0.9973	0.9943	0.0548	0.9491	4.2577
BioSyn BioBert	0.9915	0.8916	0.9784	0.9986	0.9923	0.9330	0.9955	0.9905	0.1682	0.9195	2.7164
BioSyn SapBert	0.9872	0.8904	0.9000	0.9936	0.9929	0.8952	0.9932	0.9858	0.1864	0.8756	2.5489
Bio Clinical BERT	0.9855	0.8088	0.9946	0.9996	0.9849	0.8921	0.9922	0.9837	0.2134	0.8697	2.3125
BERT	0.9828	0.9926	0.7216	0.9824	0.9997	0.8357	0.9909	0.9814	0.2568	0.8111	1.6543
BioBERT	0.9719	0.6833	0.9973	0.9998	0.9702	0.8110	0.9848	0.9687	0.1639	0.7704	2.7670
BioSentVec	0.9508	0.5516	0.9973	0.9998	0.9478	0.7103	0.9731	0.9458	0.0828	0.6458	2.5563
BioWordVec	0.9400	0.5022	0.9054	0.9936	0.9422	0.6461	0.9672	0.9344	0.1724	0.5724	1.8466
BlueBERT	0.6480	0.1464	0.9973	0.9997	0.6255	0.2553	0.7696	0.6994	0.1154	0.0653	2.7455
BioBert Lrg	0.6232	0.1383	1.0000	1.0000	0.5989	0.2430	0.7492	0.6889	0.1650	0.0491	2.0178

Table 13: Experiment 4: k-means clustering (k=3) count

Clusters	Count
0	5,730
1	254
2	133

Luke Zettlemoyer. Deep contextualized word representations. *NAACL HLT 2018*, 1:2227–2237, 2018.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

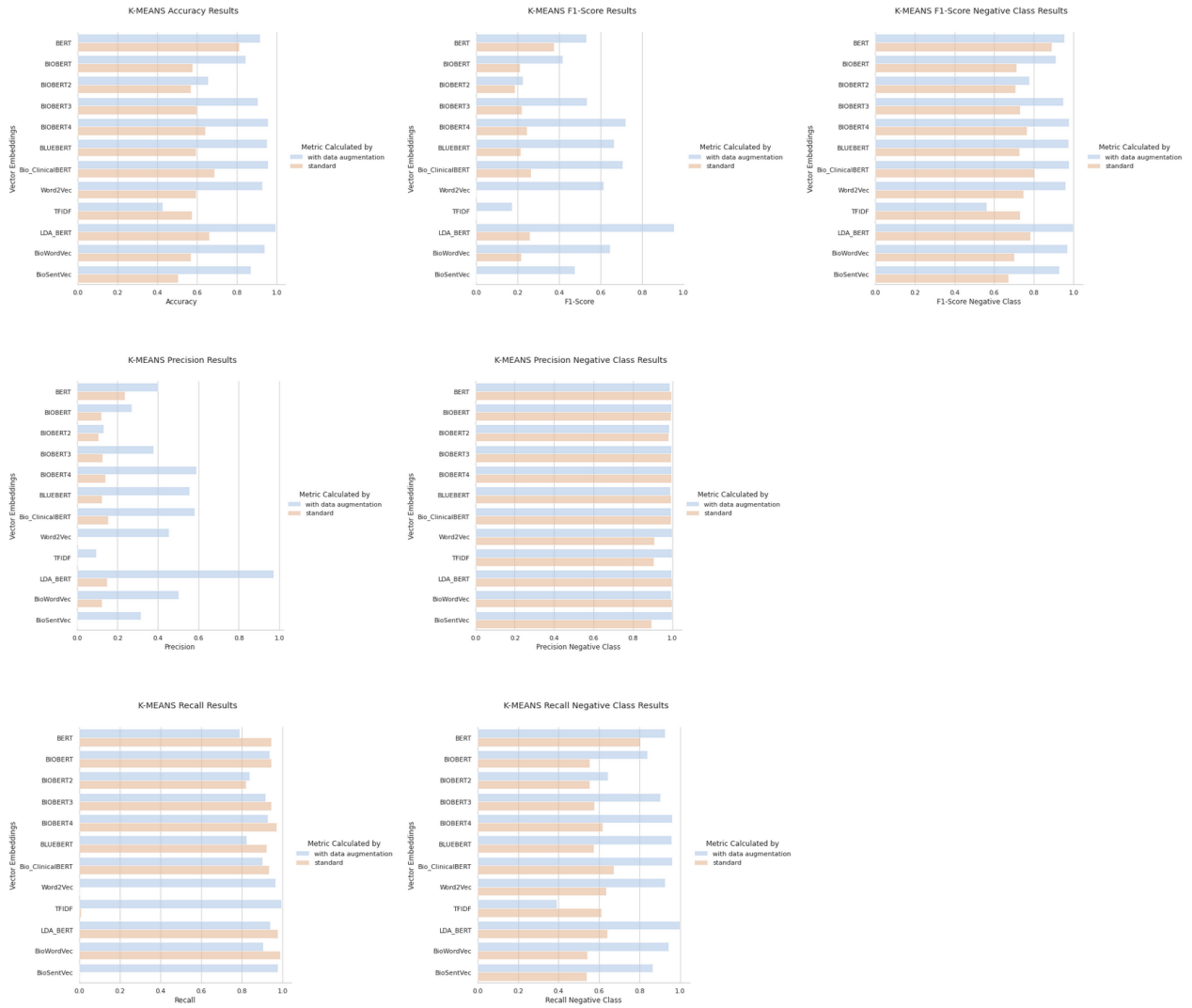

Figure 20: *Evaluation metrics k-means clustering with corrected samples by clinician, with and without data augmentation.*

Table 14: Cluster count match cluster and ground truth for Experiment 4 k-means with LDA BERT and k=3 clusters

Clusters	Ground Truth	(Clusters and Ground Truth) Count
0	0	5,722
1	1	253
2	1	109
2	0	24

Table 15: Experiment 4: GMM clustering (k=3) count

Clusters	Count
0	4,138
1	292
2	1,687

[18] S Raschka and V Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, New Haven, 3rd edition, 2019.

[19] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663, 2009.

[20] Budhaditya Saha, Sanal Lisboa, and Shameek Ghosh. Understanding patient complaint character-

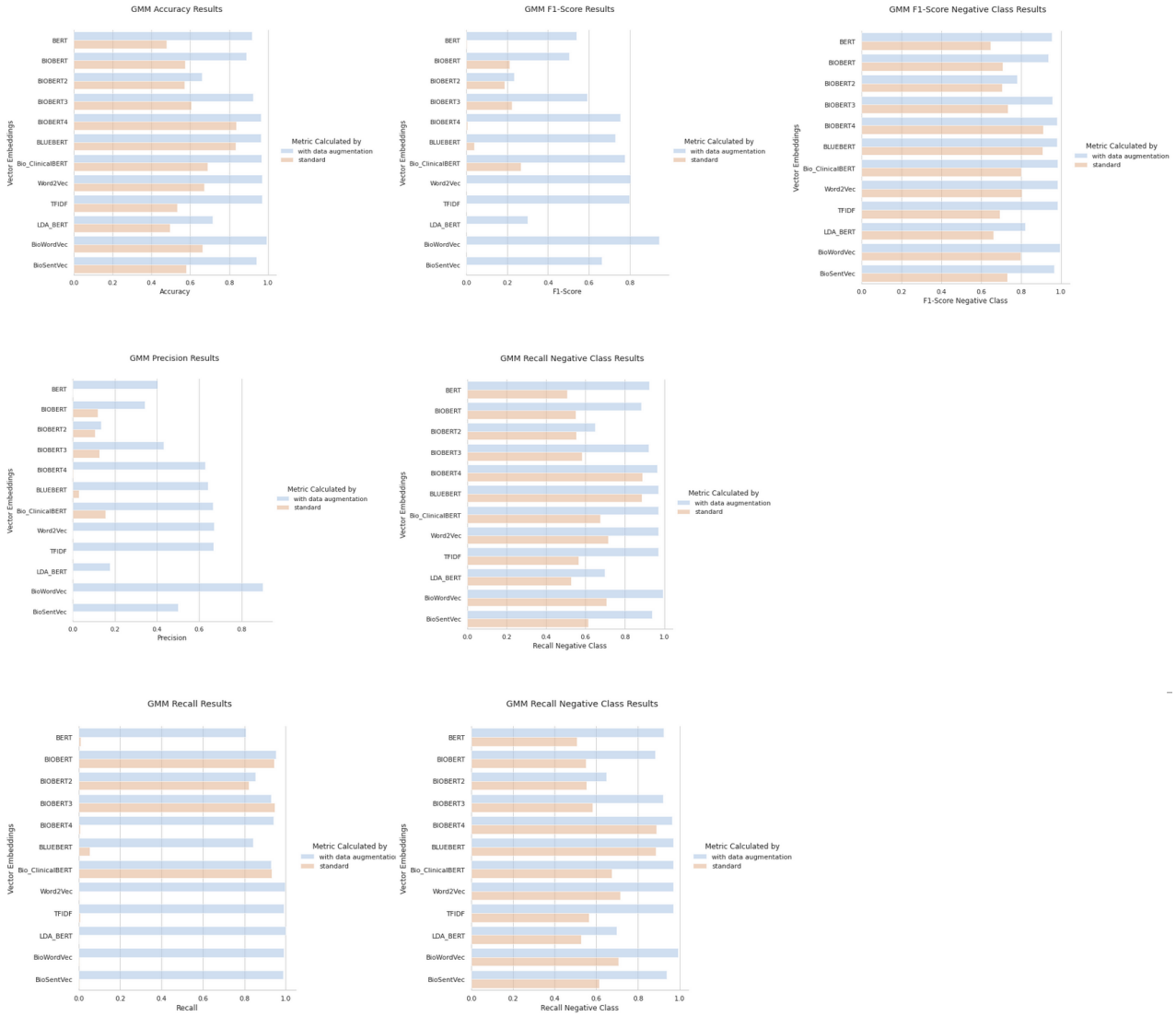

Figure 21: *Evaluation metrics Gaussian mixture model clustering with corrected samples by clinician, with and without data augmentation*

Table 16: Cluster count match cluster and ground truth for Experiment 4 GMM with BioWordVec and k=3 clusters

Clusters	Ground Truth	(Clusters and Ground Truth) Count
0	0	4,059
1	1	290
2	1	1
2	0	1,686

Table 17: Experiment 4: BIRCH clustering (k=3) count

Clusters	Count
0	5,718
1	263
2	136

istics using contextual clinical bert embeddings. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020-July:5847–5850, 7 2020.

[21] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. pages 3641–3650, 5 2020.

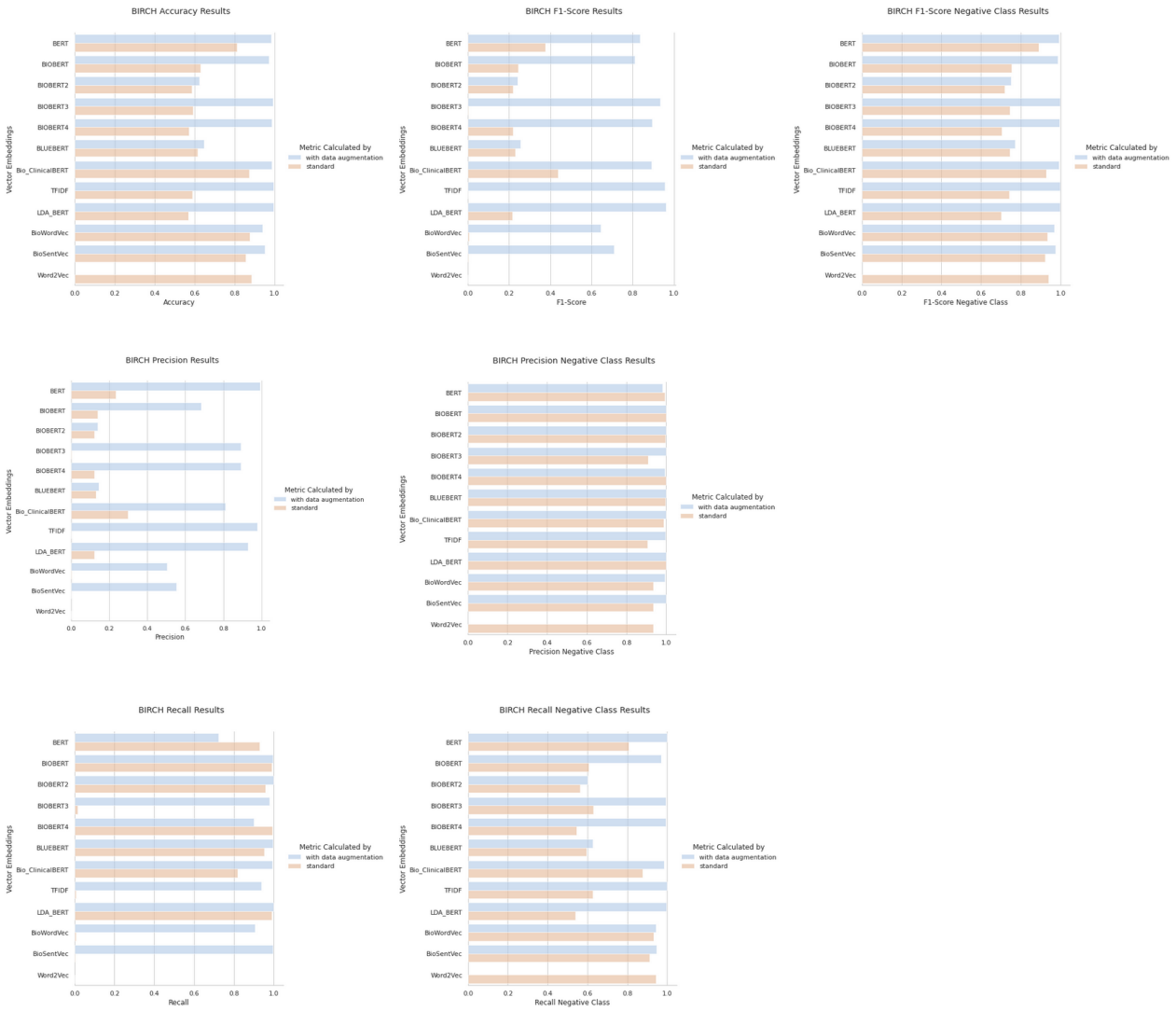

Figure 22: *Evaluation metrics of BIRCH clustering with corrected samples by clinician, with and without data augmentation.*

Table 18: Cluster count match cluster and ground truth for Experiment 4 BIRCH with LDA BERT and k=3 clusters

0	0	5,718
1	1	262
2	1	108
2	0	28

Table 19: Evaluation of results for Experiment 4

Model	Silhouette coefficient	Davies–Bouldin score
k-means LDA BERT	0.3616	1.4083
BIRCH LDA BERT	0.3353	1.3887
GMM BioWordVec	0.0841	2.3660

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5999–6009, 2017.

[23] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019*, pages 6382–6388, 2019.

Figure 23: *Word cloud results for Experiment 3, BIRCH with LDA BERT and corrected samples with data augmentation and k=2. Left (negative) and right (positive)*

[24] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch. *ACM SIGMOD Record*, 25:103–114, 6 1996.

[25] Yijia Zhang, Zhihao Chen, Qingyu; Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec,improving biomedical word embeddings with subword information and mesh. *Scientific Data 2019 6:1*, 6:1–9, 5 2019.

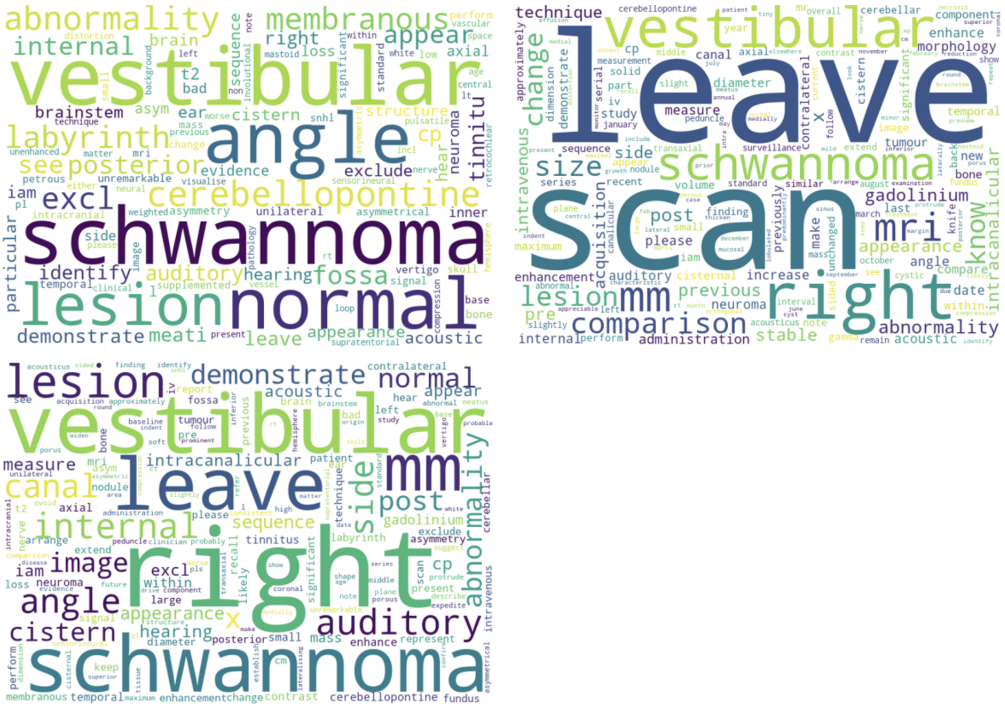

Figure 24: *Word cloud results for Experiment 4, BIRCH with LDA BERT and corrected samples with data augmentation and k=3. Negative cluster (top left), positive cluster (top right), third cluster (bottom left)*

```
MR212013688 18/07/2021 MRI Internal auditory meatus Both

Clinical History:
Exclude vestibular schwannoma. Rt asymm HL and tinnitus. Rapid drop


Findings:

There is a lesion centred within the right internal acoustic meatus
measuring up to 1.3 cm in transverse diameter, which laterally extends
towards the fundus and medially protrudes via a mildly expanded porous
acousticus into the right cerebellopontine angle.  There is no compression
of the brainstem.
Normal appearances of the left IAM.

Impression:  1.3 cm right vestibular schwannoma.

Dr ##### #####
Consultant Radiologist
GMC #####
```

Figure 25: *A sample from the third cluster containing conflicting observations*

```
Clinical History:
Left Temporal GBM, complete resection Oct 2020, previous Astrocytoma.
Please MRI 3 months, mid July 21.


Findings:

A 5mm focus of enhancement along the lateral/basal aspect of the left
temporal resection cleft is more conspicuous (series 801 image 90).
Otherwise the intracranial appearances are stable with unchanged volume of
FLAIR hyperintensity
No new intracranial diffusion restriction is identified.

No mass is identified in the brain stem, cerebellopontine angles or
internal auditory meati. No features of an vestibular schwannoma are
demonstrated.

Impression:
1. Minimal increased enhancement along the basal aspect of the left
temporal resection cleft.  Early follow-up MRI is recommended.
2. No features of a vestibular schwannoma demonstrated.

Dr ##### #####
Consultant Radiologist
GMC #####
```

Figure 26: *A sample from the third cluster indicating a pre-existing tumour in clinical history, and then a negative outcome in a later scan under findings.*

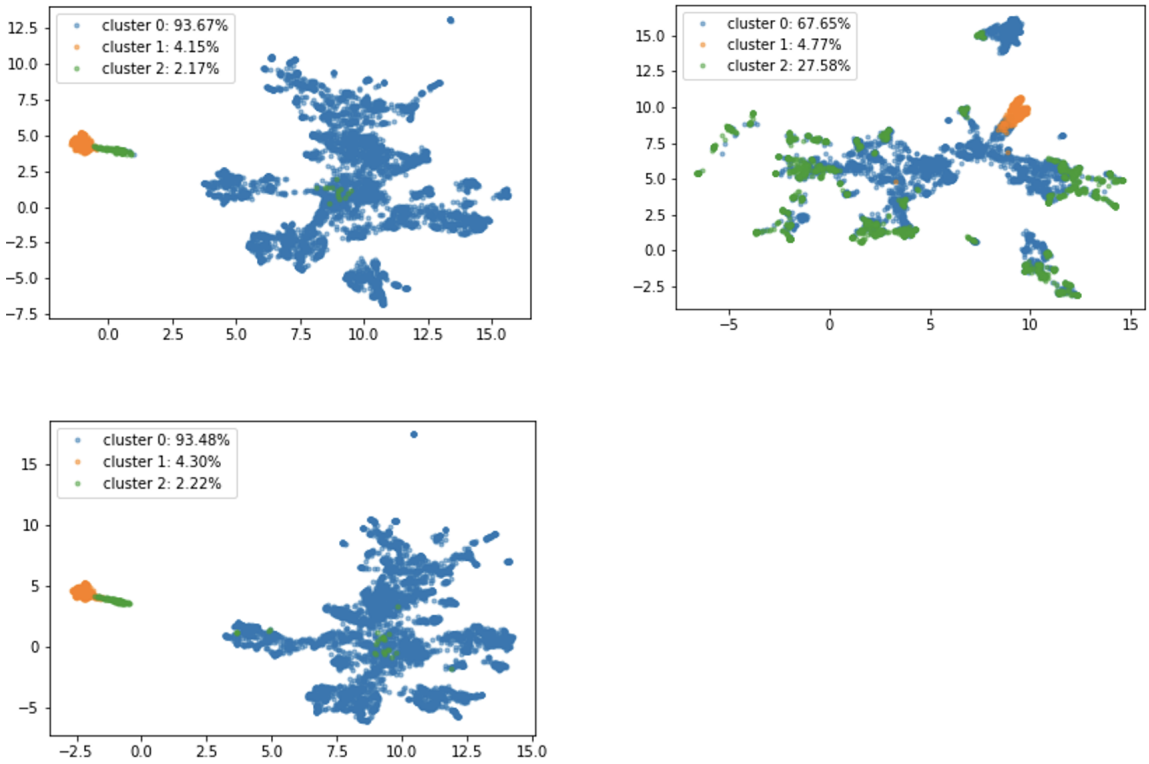

Figure 27: *UMAP visualisation of clustering results of experiment 4. Top left (k-means with LDA BERT), top right (GMM with BioWordVec) and bottom left (BIRCH with LDA BERT).*