Contents lists available online at TALENTA Publisher

DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)

Journal homepage: https://jocai.usu.ac.id

Supporting Clinical Decision Making: Semantics Based Classification of Medical Referral Letters

Ian Wilson

Computing and Mathematical Sciences, University of South Wales, Pontypridd, CF37 1DL, Wales, UK

ARTICLE INFO

Article history:
Received 15 November 2022
Revised 4 December 2022
Accepted 25 January 2023
Published online 31 January 2023

Keywords:
Clinical and Health Information Classification
Decision Systems
Natural Language Processing
Support Vector Machines
Semantic Indexing

Email:
ian.wilson@southwales.ac.uk

Corresponding Author:
Ian Wilson

ABSTRACT

This study aims to develop a Natural Language Processing based decision support system built from a repository of knowledge drawn from referral letters written between primary care doctors and specialist medical consultants. The developed system translates pre-processed referral letters into a semantic matrix of document vectors and a set of vocabulary features, based solely on the words used within each referral letter. The system applies a one-versus-rest heuristic using a Support Vector Machine (SVM) to convert a multinomial classification problem into individual binary classifications. Each document is matched to its probabilistic best fit specialism. The National Health Service Wales sourced 111,700 examples. Accuracy of 91.8% against 29 medical specialities is achieved. Accuracy increases to 97.4% and 99%, respectively, when also including one or two nearest neighbours to the best fit, providing a basis for informing the decision making of a medical professional. The study demonstrates the efficacy of using referral letters to allow or classification into specialisms and subsequent allocation of specialist care. The approach taken in this study does not require added ontologies and is readily extendable. The system offers support to medical professionals, particularly within training scenarios or where access to opinion may be in short supply.

IEEE style in citing this article:
I. Wilson, " Supporting Clinical Decision Making: Semantics Based Classification of Medical Referral Letters," *Data Science: Journal of Computing and Applied Informatics (JoCAI)*, Vol. 7, No. 1, pp. 24-34, 2023.

1. Introduction

Classifying documents based on their semantic similarity has many practical applications. One of these is the potential use of medical documentation to help inform decision making and supply a resource for areas where expertise is in short supply, such as in developing countries. The success of the contextual exploration is reliant upon the semantic matrix and design decisions. This paper reports on the creation of a semantic representation as a means for associating a given medical referral letter, written by a General Practitioner (GP), with a particular sub-specialism based solely on the words used to describe the patient's problem. Work carried out by Todd et al. [1] using internal hospital referrals shows the potential for building on referral letters. Similarly, Spasic and Button [2] performs topic modelling around patient triage but includes an ontology to achieve their outcome.

A general practitioner in Wales will write a letter to a specialist via an electronic clinical portal alongside patient biometrics and medical history. The level of detail contained within these letters is important for ensuring proper patient care as the consultant has no direct contact with the patient prior to the referral. The referral letter needs to address any questions that the consultant may have regarding the case, specifically:

- Are the symptoms correct?
- Is the priority assigned correctly?
- Is the patient being referred to the right specialist?

Should the specialist consultant have any concerns about the above three questions, the referral may be returned to the general practitioner for more clarification. This may be as simple as confirming a change in priority by the

e- ISSN: 2580-829X | p- ISSN: 2580-6769

consultant or the request for more tests to be carried out (such as an electrocardiogram) before the consultant will agree that there is a need to see the patient in a specialist clinic. The time spent adjusting the referral conditions impacts starting patient treatment and increases the workload for the general practitioner, consultant and administrators.

Whereas it is fair that document classification, particularly in the case of large datasets, is a well-developed field, this paper reports on how this was achieved without the use of an ontology and, importantly, how inclusion of the two nearest neighbours can aid with directing the patient to the right specialist. Key to this is the use of referral letters as a shared repository of knowledge that can assist primary care providers with deciding between courses of action by presenting them with alternatives. This is a significant finding given that any reduction in initially misdirected referrals will directly affect the health and well-being of patients. Importantly, partners within the health service (Digital Health and Care Wales - Research and Innovation Working Group) played a full part in defining goals and being part of the research supervision team.

Here, the authors first describe how to represent the semantics of referral letters including the transition from a set of documents to a series of encoded feature vectors. Then, the standard support vector machine method is outlined when considering both binary and multi-class problems. Next, the authors report on the generated classification model and discuss the test results found. Finally, conclusions are drawn, and future work outlined.

2. Semantic Representation

The premise of this paper is based on the notion of Statistical Semantics [3,4], where an assumption is made that "a word is characterized by the company it keeps" [5]. This is embodied by the Distributional Hypothesis such that words that occur in similar contexts tend to have similar meanings [6,7]. Healthcare records are no exception to this rule. Clinical documents feature relationships that have been mapped and reported on in literature including (but not limited to) diseases and their associated symptoms [8,9] as well as medication/testing procedures [10]. It has been shown that knowledge of pre-existing relationships within clinical data can be exploited to significant effect with techniques such as classification.

Medical document classification research focusses on these relationships within existing ontologies including the Unified Medical Language System [11] and SNOMED Clinical Terms [12] to extract specific medical terminology markers. Previous medical classification work carried out using additional ontologies can vary from monitoring patient medication [10], patient phenotyping [13] and speciality extraction [14]. Work carried out by Faris et al. [15] implements a similar method to the one discussed in this paper by using support vector machines and binary particle swarms to extract specialities from a question answering system. However, the data used in the study revolves around a small sample of short questions asked by members of the public instead of full-bodied formal letters from medical practitioners like those discussed in this report.

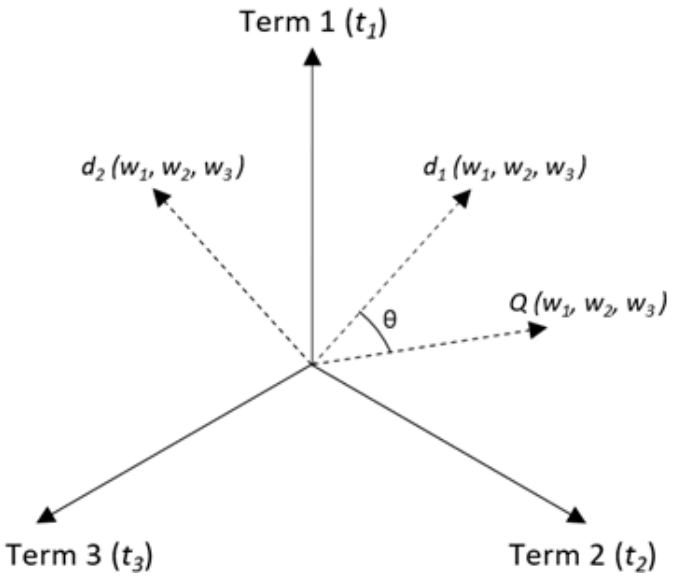

Figure 1. Vector Space Model for three terms

Classification algorithms like a support vector machine require structured data to function correctly. Attempting to parse an unstructured series of text excerpts such as the letters used in this study would yield little

to no benefit. To correct this, the approach utilises a semantic matrix as the basis for computing the distance of relatedness of documents. The matrix illustrates the occurrences of features (individual words and phrases) within the dataset separated out into individual document vectors.

The way in which the appearance of a feature is recorded will vary on a chosen encoding technique. Each feature within a document vector may occur as a frequency count (bag-of-words [16]), binary 1 or 0 (dummy encoding [17]) or a weighted value between 0 and 1 (term frequency-inverse document frequency (TF-IDF [18]). Each of these approaches to feature generation were tested using the project's dataset alongside employing the word embedding models: Word2Vec [19] and Doc2Vec [20]. Experiments found that the TF-IDF encoding method outperformed other methods [21], producing the best results and forming the basis for the following discussion. Transformer based architectures produced results comparable with TF-IDF but without advantages of efficiency (summarised as follows: BoW – 91.039, TF-IDF – 91.854, Doc2vec – 0.25, Transformer 93.1) [22].

Applying TF-IDF creates a semantic matrix of features ranked by importance [23]. The algorithm calculates the value for each word as follows. The term frequency (TF) simply ranks features (*t*) by occurrence within a single document (*d*) over the total number of words in that same document. The inverse-document frequency (IDF) deals with discerning the importance of a found term across the complete set of documents (*D*). The result is a reduction in weight for vocabulary terms that appear frequently but contain little to no importance.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (1)$$

Where:

- *t* is a vocabulary feature
- *d* is a single document
- *D* is the set of all documents
- *q* is a query document
- *tf* is the term frequency - $tf(t, d) = \frac{f_d(t)}{|d|}$
- $f_d(t)$ is the frequency of term *t* in document *d*
- *idf* is the inverse-document frequency - $idf(t, D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$
- *w* is the weight value of a vocabulary feature after encoding

Each document can be described as a vector such that $d = (w_1, w_2 \ldots w_n)$ wherein *w* is the now encoded weight for each term [24]. Figure 1 provides an example of a vector space model based on a document set with a feature vocabulary size of 3. As shown, the combined weighted value of each vocabulary term within a document vector determines its final position in vector space. The similarity of two document vectors can be calculated using a similarity measure such as cosine where the difference between two vectors is equal to the cosine of the angle between them. Calculating the similarity between a query document (*q*) and an existing document (*d)* is explained as follows:

$$similarity\ (q, d) = cos\ \theta = \frac{\vec{q} \cdot \vec{d}}{||\vec{q}||\,||\vec{d}||} = \frac{\sum_{i=1}^{n} w_{q,t} w_{d,t}}{\sqrt{\sum_{i=1}^{n} w_{q,t}^2}\sqrt{\sum_{j=1}^{n} w_{d,t}^2}} \quad (2)$$

wherein the similarity is equal to the dot product of their two unit vectors $(\vec{q}, \vec{d})$ [25]. To create comparable unit vectors, the initial vectors require normalisation to reduce the bias towards documents of longer sizes.

$$\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2} \quad (3)$$

This normalisation factor is the denominator in equation (3) where *n* is equal to the length of the document vector [26]. The positions of different documents within vector space allow for differentiation between groups of documents using methods as described below in section 3.

3. Support vector machine-based classification

The research goal here is to see whether sensible classifications could be derived solely from the semantic representations of the dataset. A well-established Support Vector Machine (SVM) classification model is utilised because of its applicability to vectors [1] of features like those derived by the TF-IDF algorithm for each document. SVMs are well documented for being the most accurate linear models used with classification problems [27,28,29], which scale well to increasingly large feature sizes, and have been used across a large range of application areas in many different fields including healthcare analysis [30]. The section first outlines the SVM training process and then describes the interaction with a multiclass dataset.

Described in 1992 by Vapnik [31], the SVM algorithm utilises hyperplanes to find the optimal decision boundary between two classes of objects. The difference between an SVM and other linear classifiers lies in the keyword optimal. An SVM focuses on the outliers of any specific class (depicted with a thin dotted line) and attempts to maximise the distance (margin) between them (2 * ***M****). Given this, for a labelled dataset with *n* examples then the input is a series of feature vectors x_i and class labels y_i:

$$(x_1, y_1) \dots (x_n, y_n)$$
$$where = \begin{cases} y_i = 1 & if\ x_i \in class\ A \\ y_i = -1 & if\ x_i \in class\ B \end{cases} \quad (4)$$

The optimal hyperplane can be defined by applying the decision function:

$$D(x) = \omega \cdot \varphi x + b \quad (5)$$

wherein ω is the weight vector, φx is the input vector and *b* is the bias. The training cycle of an SVM consists of adjusting the weight vector and bias until the optimal decision boundary is found. This allows us to then predict unknown testing data according to the following rule:

$$if\ D(x) > 0\ x \in A\ else\ x \in B \quad (6)$$

A regularisation parameter exists within SVM models denoted as *C*, in which changes in value will affect the number of misclassified points on the hyperplane. Increasing this value will result in a smaller margined hyperplane, fitting closer to overall training set and reducing the number of misclassifications. However noisy datapoints may result in hyperplanes that struggle to fit in the testing set. Conversely, reducing the size of the regularization parameter will result in larger margined hyperplanes and allow for more misclassifications during training. This parameter has more impact in non-linear classification so the default value of 1 is utilised here.

For a multi-class classification problem like natural language processing of clinical records, the SVM model can be extended to work as a one-vs-rest/one-vs-all heuristic method. This approach works by splitting down the problem into *n* binary classification problems wherein *n* is the number of different class labels. The resulting classification is then determined by probabilistic scoring wherein a document will be assigned the class label with the highest confidence prediction score.

4. Research Methods

All experiments were implemented using the PyCharm Professional Edition IDE (Python 3.7) running on Windows 10 with an AMD Ryzen 5900x and 32gb of DDR4 random-access memory (RAM). The Python libraries needed to replicate the results below are pandas [32], Natural Language Toolkit (NLTK) [33] and scikit-learn [34]. The NLTK library provides the base collection of stop words for the English language and the sci-kit learn library provides implementations of the TF-IDF encoder, LinearSVC and the performance metrics. The experimental dataset consisted of 120,572 referral letters from multiple general practitioners (GP) across Wales to hospital consultants [35].

Initial data pre-processing of the documents reduced this number to 111,700 by removing test documents and letters that fell under specialties with insufficient supporting documents. Due to the nature of the letters sourced from different health boards, further processing had to be taken into consideration to group the specialties into common groups to reduce the number of classes.

The final classes trained by the system are shown in Table 1. Data pre-processing was carried out using regular expressions to clean the data of symbols, punctuation, and capitalisations. To preserve contextual markers and to keep medical terms present in full, no use of stemming or lemmatising was carried out on the dataset. The final step involved extending the stop word list to include words and phrases unrelated to a patient's condition such as the salutation and valediction present within each letter.

Table 1. Medical specialties used for classification

Specialty	Combined Specialties	Docs
Cardiology	Cardiology (card)	3606
Care of the elderly	Care of the elderly usc ident	275
Clinical Immunology		444
Clinical neurophysiology		225
Community orthopaedic		2047
Dermatology	Dermatology (derm), Dermatology (usc), Dermatology laser, Dermatology usc ident	14760

Table 2. Medical specialties used for classification (continued)

Dietetics	Dietetics (dthe)	1682
Endocrinology	Medical endocrinology (mendoc), Endocrinology usc identifier, Endocrinology usc ident	989
Ent	Ent (usc), Ent audiological medicine (entam), Ent usc ident, Ear nose and throat (ent)	12529
Gastroenterology	Gastroenterology (gastro), Gastroenterology (usc)	5210
General medicine	General medicine (genmed), General medicine usc identifier, General medicine nurses (gmedn)	1098
General surgery	General surgery (surg), General surgery usc identifier, General surgery breast clinic (surb/c), General surgery breast service, Breast (usc), Breast, Gs breast usc	14312
Geriatric medicine	Geriatric medicine pathy day hosp (gerijp), Geriatric medicine (geri)	406
Gynaecology	Gynaecology (gynae), Gynaecology (usc), Gynaecology usc identifer	10182
Haematology (clinical)	Haematology (clinical) usc ide, Haematology-clinical (haem)	720
Nephrology	Nephrology (neph)	436
Neurology	Neurology (neur), Neurology epilepsy (epilep), Other neurology, Other neurology usc ident	2172
Oral/maxilla facial surgery	OMF usc identifier, Oral/maxilla-facial surgery (oral), Omf usc identifier	1039
Ophthalmology	Ophthalmology usc identifer	719
Orthopaedic	Orth foot & ankle (t/ofa), Orthopaedic hand (t/hand), Orthopaedic hip (t/ohip), Orthopaedic knee (t/knee), Orthopaedic paediatrics (t/paed), Orthopaedic shoulder (t/osh), Otrhopaedic spinal (t/osp), Orthopaedic spines, Orthopaedic spines usc ident, Trauma & orthopaedic, Trauma & orthopaedic usc ident, Trauma & orthopaedics (t/o)	11408
Paediatrics	Paediatric endocrine (pendo), Paediatric gastroenterology (pgast), Paediatric respiratory (presp), Paediatric cardiology (pcardl), Childrens ent (entpae), Paediatrics usc identifier, Paediatric surgery (paedsu)	3646
Pain management	Chronic pain management(chronp)	807
Physiotherapy adult		7072
Rapid diagnostic centre	Rapid diagnostic centre usc	152
Rehabilitation	Rehab day hospital (rehadh), Rehabilitation (rehab)	656
Rheumatology	Rheumatology (rheum)	2141
Thoracic Medicine	Thoracic medicine (throme), Thoracic medicine usc ident, Respiratory (usc)	2912
Urology	Urology (usc), Urology (urol), Urology usc ident	8706
Vascular surgery	Vascular	777

For the classification work carried out in the experiment, other algorithms for classification were tested including Bayesian classifiers, random forest decision trees and logistic regression. Table 2 shows the average F1-Score for the six models tested on the dataset. These averages are the result of performing 5-fold cross validation on the dataset and show that with when using a TF-IDF vectorisation technique, the support vector machine has a significantly higher F1-score than the other five models. As a result, the SVM is the algorithm chosen for the experiments and associated outcomes in the rest of the paper.

Table 3. F1-Scores for classifying 6 models to the dataset using TF-IDF matrices

Model	**F1-Score**
Linear Support Vector Machine	91.914
Logistic Regression	88.260
Stochastic Gradient Descent	85.690
Random Forest	81.303
Bernoulli Naïve-Bayes	73.226
Multinomial Naïve-Bayes	73.331

4.1. Classification

The goal of document classification is to accurately assign each document to an associated label. The method implemented within here focusses solely on the already existing information within each of the referral letters and no outside ontology. Metrics for determining classification accuracy are displayed in Table 3 using formulas taken from Sokolova and Lapalme [36]. The formulas used for these metrics are specific to representing micro-averaging within multiclass classification tasks, which provides the means to ascertain the ability of the linear support vector machine when carrying out a classification task. F-Measure combine precision and recall into a single metric. β is used to indicate the changing variable within the F-measure. Changes to this variable allow for higher weighting to be placed on the precision or recall values. This project uses the F-Measure with a β value of 1 (F_1-Score) as precision and recall are equally important.

Table 4 Performance Metrics for Document Classification

Performance Metric	Formula		Meaning
TP - True Positive	n/a		Number of documents correctly labelled as class *i*
TN – True Negative	n/a		Number of documents correctly not labelled as class *i*
FP – False Positive	n/a		Number of documents incorrectly labelled as class *i*
FN – False Negative	n/a		Number of documents incorrectly not labelled as class *i*
Precision	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	(1)	True positives over all documents labelled as positive for that class.
Recall (Sensitivity)	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}\left(tp_i+f_{n_i}\right)}$	(2)	True positives over all documents that belonged to that class.
F-Measure	$\frac{(\beta^2+1)Precision_\mu Recall_\mu}{\beta^2 Precision_\mu+Recall_\mu}$	(3)	Micro averaged harmonic mean between precision and recall.

Section 2 discussed the transformation of referral letters into a semantic matrix using the tf-idf algorithm. The referral letters are written in full and not as a series of short medical terms, it is necessary to decide on a list of features to use within the dataset. Steps were taken to reduce the size of the vocabulary by implementing a minimum and maximum document frequency. The minimum frequency was set to 5 so that a feature needed to occur in at least 5 documents across the set. The maximum was set to 0.1, which results in any feature occurring in more than 10% of all documents to be considered unimportant and excluded. The range of n-grams captured as features was set between 1 and 3. This has been done as medical terms are known to exist in more than one-word phrases, and the surrounding context of a word can drastically change its meaning. The Linear SVM used implements the one vs rest heuristic method for multiclass classification with a value of 1 for the regularisation parameter, as discussed in section 3.

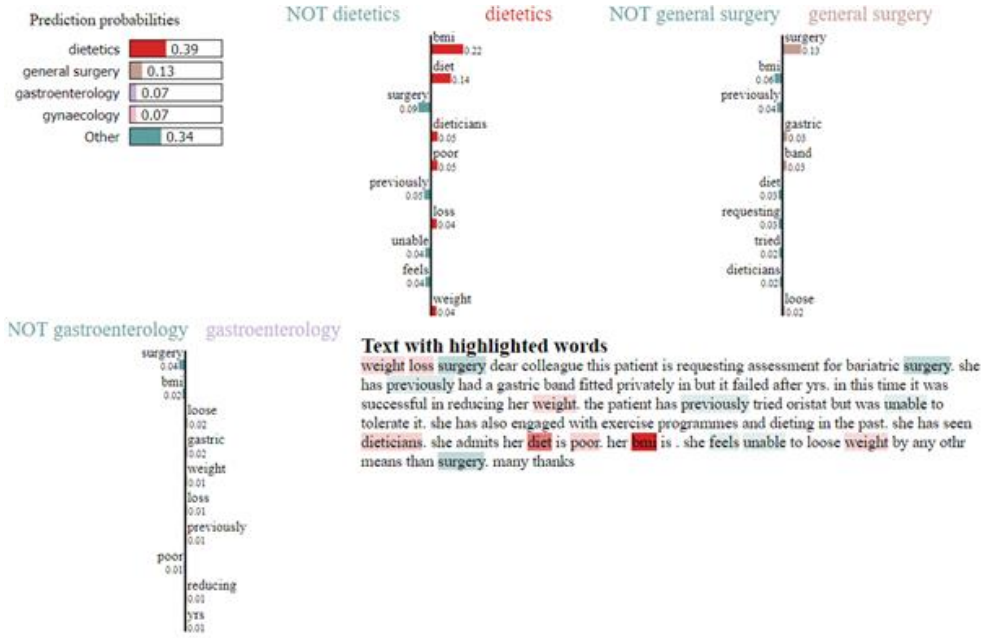

Figure 1 Lime explanation for classifying referral letter to specialties.

With the goal of the model being a support tool instead of a medical practitioner replacement, work has been carried out using a second, logistic regression-based classifier on top of the SVM to curate a matrix of probabilities related to the outcomes of the model used. Manual inspection was carried out on elements to determine if the errors occurred when multiple class were predicted with high probability or if one class was outweighing others significantly. Figure 2 provides an insight into how these results could be presented to a medical practitioner. Here, the python library lime [37] has been used to show an example of an initially misdirected referral letter.

5. Result

The results in Table 4 show the individual precision, recall and F_1-Score for each class alongside accuracies for the overall dataset. The weighted average F_1-score accounts for the class imbalance between the different labels. The results show that while the model achieves an accuracy of 92% across the data, there are a few classes wherein the model is unable to separate between the close links (orthopaedic and community orthopaedic as an example). Whilst these specialties deal with similar issues, the nature of what medical professional deals with and how the patients are treated afterwards in hospital requires them to be kept separated. However, as the goal of the research is to support medical practitioners, returning a short-list of potential specialties that a document belongs to alongside the reasonings for doing so is more beneficial and trustworthy than just a single outcome displayed to the user.

Table 5 Testing set accuracies of the LinearSVC whilst using TF-IDF vectorisation.

	precision	**recall**	**f1-score**	**support**
Cardiology	0.95	0.97	0.96	682
Care of the elderly	0.87	0.66	0.75	59
Clinical immunology	0.9	0.86	0.88	88
Clinical neuro-physiology	0.89	0.73	0.8	44
Community orthopaedic	0.73	0.52	0.6	393
Dermatology	0.97	0.97	0.97	3017
Dietetics	0.96	0.92	0.94	346
Endocrinology	0.85	0.82	0.83	201
ENT	0.95	0.97	0.96	2551
Gastroenterology	0.84	0.88	0.86	987
General Medicine	0.83	0.67	0.74	202
General Surgery	0.93	0.93	0.93	2875
Haematology	0.9	0.88	0.89	146
Nephrology	0.94	0.89	0.91	87
Ophthalmology	0.87	0.88	0.87	126
Oral/Maxillo facial surgery	0.9	0.79	0.84	218
Orthopaedic	0.86	0.92	0.89	2327
Paediatrics	0.9	0.77	0.83	741
Pain Management	0.85	0.76	0.8	147
Physiotherapy	0.85	0.86	0.86	1348
Rapid diagnostic centre	0.91	0.32	0.48	31
Rehabilitation	0.92	0.88	0.9	128
Rheumatology	0.91	0.91	0.91	430
Thoracic medicine	0.94	0.97	0.96	594

Table 6 Testing set accuracies of the LinearSVC whilst using TF-IDF vectorisation (continued).

Urology	0.95	0.98	0.96	1766
Vascular surgery	0.79	0.84	0.81	146
Gynaecology	0.97	0.97	0.97	2024
Neurology	0.91	0.86	0.89	441
Geriatric medicine	0.88	0.78	0.82	81
accuracy			**0.92**	**22226**
macro average	**0.89**	**0.83**	**0.86**	**22226**
weighted average	**0.92**	**0.92**	**0.92**	**22226**

6. Discussion

Upon applying a calibration classifier as a mask over the initial linear support vector machine, it was found that for a large number of cases present within the dataset, a misclassification may be the result of the model failing to decipher the difference between two to three specialties. After extending the outcomes to include the top two probabilities, the accuracy rose from 91.8% (20402 documents) to 97.4% (21661 documents). Further extension to include the third outcome resulted in an accuracy value of 99% (21964 documents).

Figure 2 provides an example of the output from the python library lime [37] showing how a clinician's referral letter has been classified by a machine learning model, which helps to explain reasoning behind the model's choice of outcome. The output shows how the model has classified the data and that key indicators such as diet and bmi have the greatest influence on deciding that this particular letter belongs to dietetics. It also shows that other elements in the letter such as the word surgery and gastric band influence the model towards a different specialty (general surgery). The referral letter displayed was assigned to the general surgery specialty by the general practitioner. Had the general practitioner had access to alternatives suggested by their peers for similar cases then they may have made a different decision.

A manual inspection of cases where the system's suggested specialism did not align with the actual referral letter included symptoms that suggested a different specialty to the one assigned by the medical professional. Document classification is and will continue to be an essential part of extending the usability of electronic health systems. Numerous documents are passed between personnel via clinical portals each day. The use of text-only fields within the letters can lead to missing or misinterpreted information within communications between GP and hospital consultant resulting in an avoidable iteration of the referral process.

The inclusion of the proposed system within processes adopted by NHS Wales or the wider medical community could reduce the amount of resources needed by hospital consultants and GPs when referring patients for specialist treatment. However, it is acknowledged that the scope of semantic context for any feature only extends to the words and phrases captured by the TF-IDF vectorizer, which may result in the loss of contextual information held elsewhere in the sentence.

6.1. Future work

The next goal is to produce an interface for medical professionals without the need of including an external visualisation library such as lime to decipher the outcomes of the models, which can then go to clinical trial. In parallel, the work will be extended to include pretrained Deep Learning models such as Google's BERT. Also, the inclusion of named entity recognition (NER) [38] can be included in the pipeline to reduce the overall size of the vocabulary by grouping variations of the same feature together [39].

7. Conclusion

Here, the authors reported on the application of a well-established and understood classification method to identify specialties within a previously unexplored dataset consisting of NHS Wales GP referral letters. The approach taken is explained in detail to help others apply the approach to a similar dataset. The reported method does not require the inclusion of ontologies and, as such, has the advantage of relying solely upon the documents themselves (see Section 4.1) whilst minimising the system architecture size. The paper addresses the means to translate plain text into a semantic matrix comprising of document vectors and an associated vocabulary of features. The proposed system employs a one-versus-rest heuristic strategy using an SVM to convert a multi-class classification problem into several individual binary classifications. The system's first choice aligned with that of the medical professional in 91.8% of cases, which increased to 99% with the inclusion of the two nearest neighbours. As the goal here is to

provide a decision support system and in no way replace the medical professional, the system's capacity for drawing on and presenting comparable outcomes to the user strongly suggests that this type of system can help inform decision making and contribute to the overall efficacy of the medical pipeline.

Declarations

Research Funding

This work is support by KESS under grant number 21082. Knowledge Economy Skills Scholarships (KESS) is a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys. Publication costs are covered using allocated KESS funding.

Disclosure statement

The authors have no financial or proprietary interests in any material discussed in this article.

Data availability statement

The authors acknowledge the contribution of Digital Health and Care Wales (formerly National Health Service Wales Informatics Service) in sourcing the data for this research in compliance with the Open Government Licence [40], for their support in securing the research funding, setting of objectives and subsequent contributions to its supervision and direction.

A subset of the dataset supporting the conclusions of this article has been anonymised and is available in the University of South Wales Intelligence Research repository, (Pwd: ESSSbMRLuaSVM) https://intelligence.research.southwales.ac.uk/documents/3573/Letters.zip.

Author contributions

Conceptualization: Laurence Jones, Ian Wilson; Data curation: Laurence Jones, Ian Wilson; Formal analysis: Laurence Jones, Ian Wilson; Funding acquisition: Ian Wilson; Investigation: Laurence Jones; Methodology: Laurence Jones; Project administration: Ian Wilson; Resources: Ian Wilson; Software: Laurence Jones; Supervision: Ian Wilson; Validation: Ian Wilson; Visualization: Laurence Jones, Ian Wilson; Roles/Writing - original draft: Laurence Jones; Writing - review & editing: Ian Wilson

References

[1] J. Todd, B. Richards, B. J. Vanstone and A. Gepp, "Text Mining and Automation for Processing of Patient Referrals," *Applied Clinical Information*, vol. 9, no. 1, pp. 232-237, 2018.

[2] I. Spasic and K. Button, "Patient Triage by Topic Modeling of Referral Letters: Feasibility Study," *JMIR Medical Informatics*, vol. 8, no. 11, 2020.

[3] G. W. Furnas, T. K. Landauer, T. K. Gomez and S. T. Dumais, "Human factors and behavioural science: Statistical semantics: Analysis of the potential performance of keyword information systems.," *Bell System Technical Journal*, vol. 62, no. 6, pp. 1753-1806, 1983.

[4] W. Weaver, Machine Translation of Languages, W. Locke and D. Booth, Eds., Cambridge, Massachusetts: MIT Press, 1955, pp. 15-23.

[5] J. R. Firth, A Synopsis of Linguistic Theory 1930-1955 'Studies in Linguistic Analysis', Oxford, 1957.

[6] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2-3, pp. 146-162, 1954.

[7] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, Raleigh, North Carolina, 2010.

[8] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR Med Inform*, vol. 7, no. 2, 2019.

[9] X. Wang, A. Chused, N. Elhadad, C. Friedman and M. Markatou, "Automated Knowledge Acquisition from Clinical Narrative Reports," in *AMIA Annu Symp Proc.*, Washington, DC, 2008.

[10] Ö. Uzuner, I. Solti and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514-518, 2010.

[11] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. 1, pp. D267-D270, 2004.

[12] M. Q. Stearns, C. Price, K. A. Spackman and A. Y. Yang, "SNOMED clinical terms: overview of the development process and project status.," in *AMIA Symposium*, 2001.

[13] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Cralson, J. T. Wu, J. Welt, J. Foote Jr., E. T. Moseley, D. W. Grand, P. D. Tyler and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS ONE* , vol. 13, no. 2, 2018.

[14] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, 2017.

[15] H. Faris, M. Habib, M. Faris, M. Alomari and A. Alomari, "Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines," *Journal of Biomedical Informatics*, 2020.

[16] Z. S. Harris, "Distributional Structure," *Word*, vol. 10, no. 2, pp. 146-162, 1954.

[17] G. Boole, An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities, London: Cambridge: Macmillan and Co., 1854.

[18] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval.," *Journal of Documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[19] T. Mikolov, K. Chen, G. Corrado & J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv. https://doi.org/10.48550/arXiv.1301.3781

[20] Quoc Le, Tomas Mikolov Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, 2014.

[21] J. Devlin, M. Chang, K. Lee & K. Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. https://doi.org/10.48550/arXiv.1810.04805

[22] L. Jones 2023, 'Natural Language Processing as a Toolin Supporting Clinical Decision-Making', PhD thesis, University of South Wales, UK.

[23] G. Salton and M. J. McGill, Introduction to Modern Information Retreival, McGraw-Hill, Inc., 1986.

[24] D. L. Lee, H. Chang and K. Seamons, "Document ranking and the vector-space model," *IEEE Software*, vol. 14, no. 2, pp. 67-75, 1997.

[25] V. N. Gudivada and C. R. Rao, Computational analysis and understanding of natural languages, Amsterdam: Elsevier, 2018.

[26] A. Bagga and B. Baldwin, "Entity-based cross-document coreferencing using the Vector Space Model," in *ACL '98/COLING '98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1998.

[27] S. Dumais, J. Platt, D. Heckerman and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, New York, 1998.

[28] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, 1999.

[29] D. Meyer, F. Leisch and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1-2, pp. 169-186, 2003.

[30] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[31] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, New York, 1992.

[32] W. McKinney and others, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010.

[33] E. Loper and S. Bird, "NLTK: the Natural Language Toolkit," in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, Stroudsburg, 2002.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pasoss, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine learning Research*, vol. 12, pp. 2825-2830, 2011.

[35] Digital Health and Care Wales, *Doctor referral letter dataset*, 2019.

[36] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.

[37] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should {I} Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining*, San Francisco, ACM, 2016, pp. 1135-1144.

[38] I. Augenstein, L. Derczynski and K. Bontcheva, "Generalisation in named entity recognition: A quantitative analysis," *Computer Speech & Language*, vol. 44, pp. 61-83, 2017.

[39] G. K. Savova, J. J. Masanz, V. P. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the Americ*, vol. 17, no. 5, pp. 507-513, 2010.

[40] UK Government, "Open government licence for public sector information," 2014. [Online]. Available: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. [Accessed 25 March 2021].