Healthcare Analytics 5 (2024) 100288

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews

Corban Allenbrand

The University of Kansas, School of Business Department of Analytics, Information, and Operations Management, 1654 Naismith Drive, Lawrence, KS 66049, United States of America

ARTICLE INFO

Keywords:
Supervised machine learning
Unsupervised machine learning
Topic models
Latent dirichlet allocation
Cluster analysis
Non-linear support vector machine

ABSTRACT

Optimization of medication therapy depends on maximizing benefits and minimizing side effects of medications. This research showed how a joint approach using text mining, natural language processing, and machine learning can provide information for personalized and optimized medication therapy. Reviews on the benefits and side effects of prescription and over-the-counter medications were used to determine how well an integrated supervised and unsupervised learning could learn medication satisfaction. Supervised learning with naïve-Bayes, non-linear support vector machine with radial basis function kernels, and random forests with CART decision trees was measured by a micro-aggregated Matthews correlation coefficient and a macro-averaged F1 measure. Random forests outperformed support vector machines by almost 250% and naive-Bayes by 600% on the two evaluation metrics. All models did better with three rating levels, instead of five. Topic modeling and stacked cluster analysis were coupled with parts-of-speech tagging and text mining operations to establish a robust data preprocessing procedure to eliminate noisy features from the data. Unsupervised topic modeling and clustering represented an exploratory validation of how easy supervised classification would be. Well-defined latent topics were discovered including topics on “sleep quality”, “the opportunity to get back to work”, and “weight gain”. Overlapping clusters revealed that incorporating more information on social, demographic, or medical history variables could improve classifier performance. This research provided evidence that medication satisfaction can be learned with carefully designed joint supervised, unsupervised, and natural language learning techniques.

1. Introduction

Customization of goods and services has been shown to drive innovation in a wide range of areas such as manufacturing, marketing, and information technology [1,2]. A multitude of benefits emerge when personalizing a service, such as medical or healthcare, to a consumer’s demand, including enhanced usability, satisfaction, and persistence of use [3]. Health and medical care is a prime candidate for enhanced personalization. In particular, enhancements to medication therapy can yield advantages for the patient such as optimal therapeutic responses, increase safety and efficacy, decrease adverse drug reactions, enhanced patient compliance, and reduced cost [4–6]. A step beyond personalization is medical and health care that unifies systems medicine with digitalization and analytics to foster personalized, predictive, preventative, and participatory (P4) care [7–9]. A strong coupling of data science with P4 medicine offers greater precision to health and medical care, were sources of patient-specific variability are identified and mitigated to achieve better outcomes [10,11].

Success of personalized medicine is not guaranteed, and is fundamentally based on close cooperative relationships between patients and the providers of the care [12,13]. Participation by the patient in the relationship can occur at many phases of the healthcare pipeline, including the at the delivery of feedback about medication therapies. Accurate and timely medication therapy is paramount to the success of treatment. Data scientific and machine learning methods are needed to meet the patient the feedback phase where as much information from them can be passed to healthcare decision-makers [14]. An example of where a system of data science, machine learning, and P4 medicine has great potential is with medication nonadherence. Medication nonadherence is an observed difficulty of a patient to take medications as directed, which has been estimated to incur an annual cost of 100-300 billion USD [15,16]. Medication nonadherence is a clear barrier to optimized health care and necessitates innovative solutions [17]. Nonadherence is partly due to a patient’s dissatisfaction with the prescribed medication [18]. To maximize usage, it is important to understand how a person uses a particular drug, perceives its safety, recognizes its effectiveness, and responds to unintended effects of the medication [19]. Medication safety and efficacy are established for small

E-mail address: callenbrand@ku.edu.

https://doi.org/10.1016/j.health.2023.100288
Received 16 March 2023; Received in revised form 25 September 2023; Accepted 4 December 2023
Available online 6 December 2023

groups of people during clinical trials; however, these results rarely perfectly predict how any single individual person will respond to the product [20,21]. What are needed are methods to reliably extract information about medication satisfaction and preferences directly from the patient, without distortion from the influence of healthcare providers, and to pass that information on to healthcare decision makers.

Across academics and healthcare practitioners, it is generally believed that the science of data [22] alongside analytics, machine learning [23], and artificial intelligence [24,25] are uniquely positioned to best measure, analyze, and understand a patient's response to medication therapies [26]. The quality of medications is measurable in terms of two dimension, side effects and benefits. Useful information on medication therapy satisfaction and preference must correctly extract information on the two dimensions of side effects and benefits. Patient-reported outcomes in the form of written statements are a valuable source of feedback for healthcare decision makers [27]. However, feedback from patients, left unstructured, would pose considerable challenges for automated learning of patterns with most information systems available to decision-makers [28]. It is this reason that a formal method that combines text mining, natural language processing, and unsupervised and supervised machine learning could mediate the feedback channel between patient and provider.

2. Contributions

This research investigated how much information about medication preferences is contained in self-reported reviews of drug medications and if automated retrieval of that information through text mining, natural language processing, and machine learning was possible. It demonstrates a formal method that starts with raw text from a patient on the benefit and side effects of a medication and ends with predictions on the patient's preferences for that medication. A conceptually important contribution is that reviews of medications were shown to carry information on the two dimensions of medications most important to patients, benefits and side effects, and that this information offers predictive value in understanding satisfaction with the medication. The predictions summarize a patient's satisfaction and experience with a medication, thus informing institutions such as hospitals, health insurances, and drug companies on how to optimize health care delivery.

The method offers a novel combination of unsupervised hierarchical, k-means, and density-based clustering with latent Dirichlet allocation topic modeling to enhance data preprocessing and make it more suitable for supervised machine learning classifiers. Automated preprocessing with the clustering and topic modeling results was paired with part-of-speech tagging and a carefully crafted sequence of text mining operations to decrease noise in the medication reviews. Sources of noise included spelling errors, symbols, rare terms, and parts-of-speech that would otherwise impair downstream supervised learning. The joint supervised and unsupervised learning can help with the identification of which word features in the reviews most accurately reflect satisfaction levels, find associations between certain words or phrases in the reviews and satisfaction levels, categorize and rank the importance of benefits or side effects communicated in the reviews, and compile a list of words or phrases that could be used as prompts in more structured patient interviews.

3. Background and related works

Machine learning (ML) based systems for text classification, including naïve-Bayes, logistic regression classifiers, support vector machines, random forests, and neural networks have been extensively evaluated in the literature [29–32]. Advantages and disadvantages of the different ML approaches to text classification depend on characteristics of the classification problem such as the number of predicted classes, data attributes like size or feature quality, balance of the outcome classes, computational cost of evaluation, speed of model training, and availability of data for testing [33]. Sentiment analysis and opinion mining is a branch of natural language processing (NLP) that approached text classification with machine learning in order to uncover an individual's emotional, evaluative, or judgmental thoughts [34]. In many fields, including travel, dining, consumer goods, entertainment, and banking, NLP and ML have been combined to extract sentiment, opinion, and rankings from consumer written reviews [35,36].

Extraction of evaluations and sentiment from drug reviews scraped from online sources with ML methods has been attempted by others including Graber et al. [37] who used data similar to that used in this research, but employed logistic regression models trained on simple lexical features to predict sentiment of drug reviews through a cross-domain and cross-data sentiment analysis approach. Classification results were sensitive to the domain and the data used for training. Work by Shiju and He [38] also examined drug review classification with various neural network models but focused only on accuracy and the average of precision and recall to assess model performance. Both of these works neglected natural language processing aspects of the medication reviews such as clustering, topic modeling, and parts-of-speech tagging. This current research integrates supervised and unsupervised ML with topic modeling and natural language processing to synthesize a more complete analytics pipeline.

Lexicon-based approaches are common and entail mapping user text data to established vocabularies with known and labeled sentiment [39]. Supervised classification in the drug domain has received less attention than the lexicon-based approach. A rule-based system for polarity classification of drug reviews was developed in [40] where annotation of a user's review of a drug was found to be a challenging task when done manually. Manual sentiment annotation was merged with ML automated classification on manually annotated user-written web messages in [41] to understand drug facts. The quality of the sentiment labels were verified with the ML classification accuracy, and it was found that several different ML algorithms achieved decent performance on human written drug statements. Comparison of the accuracy of different supervised learning methods on the determination of drug satisfaction from written drug reviews found support for neural networks over support vector machines in terms of superior precision, recall and F1-score [42].

Text analysis and understanding extends beyond supervised classification. Topical structure of documents and other collections of discrete data can be modeled with generative probabilistic frameworks, namely, latent Dirichlet allocation (LDA). In LDA, a written document is modeled as a being generated from a latent mixture of topics and that each topic is a distribution over the words of a vocabulary [43]. Latent topic models returned from LDA can be utilized as features in standard supervised learning to enhance text categorization [44]. In addition to LDA, which can be seen as a type of probabilistic clustering algorithm, several other clustering algorithms can be used to ascertain document groupings and similarity in the context of text. Clustering algorithms must be adapted to text data. Such as careful choice of distance functions as text data has unique properties such as high dimensionality, sparsity, and non-uniformity of size [45]. Other classes of clustering algorithms have been used for text data, including hierarchical, partitioning, and density-based, all with distinct trade-offs with respect to efficiency and effectiveness [46]. Many times, a combined clustering approach is required to properly capture text-based similarity [47]. In this research, cluster analysis served as an unsupervised validator on the clustering behavior of the review documents to ascertain ease by which the reviews could be separated and classified. If coherent and well-defined topics and clusters were identifiable, then it was posited that classification would be more feasible.

3.1. Machine learning algorithms

General overviews of the three machine learning algorithms used in this research will be provided. Alternative models based on generalized linear models like logistic regression [48,49], or its penalized counterpart [50], were not investigated as they have already received much attention [51–53] and alternative classifiers have been shown to demonstrate better performance [54,55]. Models based on k-nearest neighbor were not used for classification, as cluster analysis was used to supplement the data preprocessing instead [56,57]. Models based on neural networks [58], and the transformer architecture in particular [59], are adept classifiers [60,61], but were not used because of the greater lack of explainability and interpretability with neural network based models.

3.1.1. Supervised

Naïve-Bayes is a family of conditional-probability models based on a shared principle, conditioned on a class variable, all feature values are mutually independent [62,63]. Naïve-Bayes classifiers combine the naïve-Bayes model plus a decision rule, such as maximum a posteriori, which selects the most probable class hypothesis, given the data [64]. Naïve-Bayes algorithms are computationally tractable but do have known limitations, including a strong assumption of conditional independence of n-grams in the reviews' text [65,66].

Support vector machines (SVM) are non-probabilistic linear classifiers that can achieve non-linear classification with a similarity function, known as a kernel function, which can separate document classes with a non-linear boundary [67]. Classification of the drug reviews was expected to benefit from the non-linear separation, as the hundreds of n-grams used to classify reviews likely created a feature space where the boundary between rating classes was not linear. The kernel used in this research was the Gaussian radial basis function kernel (RBF) and was employed because of its general acceptance and good performance [68]. Discovery of a separation boundary between documents in the feature space of n-grams was the goal of the SVM. Two important tuning hyperparameters, $(C > 0)$ and $\gamma > 0$, managed the learning and predictive performance of SVMs and required optimization [69]. As discriminative models, SVMs can be insensitive to complex relationships between the text features and target classes [70]. Interpretability of hyperparameters and support vectors is low, as they are not intrinsically linked to the original.

Random forests are an ensemble learning method that combine several decision trees to obtain predictive performance higher than what would be achievable with any single tree [71]. An ensemble is more flexible than single decision trees and more capable of exploration in a larger hypothesis space [72]. It was expected that complex relationships between n-grams and rating classes were more readily modeled with random forests due to an inherent diversity of individual tree learners [73]. Another strength of random forests is that variance and overfitting to training data is bounded above, so additional trees can be used to capture more of the hypothesis space [74]. Construction of a random forest requires several choices and parameters to optimize such as the base tree algorithm, number of trees, depth of trees, size of terminal nodes, and rate of data subsampling [66]. As an ensemble, random forests are not intrinsically interpretable like individual decision trees and although random forests are relatively quick to train on text data, ensembles of thousands of trees with large depths can create a time complexity challenge [33].

3.1.2. Unsupervised

Cluster analysis was concerned with the problem of grouping reviews so that similar reviews are contained in the same cluster and dissimilar reviews were in different clusters [46,75]. Reviews were defined by location vectors in the feature space of n-grams. Distance between the reviews was to be determined by a distance function such as Euclidean, taxicab, cosine similarity, Jaccard, or correlation-based distance [47,76]. It was decided that cosine similarity would be used because of its ability to adjust for the length of reviews, it is a measure of orientation similarity and ignores magnitude. Cluster approaches are broadly divided into hard and soft clustering groups. Hard clustering, which is used in this research, imposes a clear yes-or-no membership relation to a cluster, whereas soft clustering allows for probabilistic membership [45,75,76]

In a collection of reviews, it was surmised that a finite set of abstract semantic topics captured the semantic regularities in those reviews. Topic models are probabilistic generative models that infer semantic structures and their estimated topics [77]. Central to topic models is the premise that a review is a distribution over topics, and topics are distributions over words [44]. Many types and extensions of topic models are available, but latent-Dirichlet allocation (LDA) was elected as a candidate to ascertain if a latent semantic structure was discoverable [43]. Alternative topic models such as latent semantic analysis, correlated topic model, and dynamic topic models are available, but LDA was believed to offer the best combination of performance and interpretability [78]. Transformer-based topic models such as BERTopic [79,80] are also available but require considerably more data than available in this research to ensure reliable conclusions. Simulation of the generative process that gives rise to the reviews was the objective of LDA [81]. Review generation was assumed to proceed as follows: for every word in a review, determine a topic assignment for that word, and select a word from this topic. Hence, each review was a random mixture over latent topics and each topic is characterized by a distribution over words. Topics discovered were thought of as word clusters whose linkage is one of semantic similarity [82]. Clearly demarcated topics would raise confidence that text classification with machine learning models would be well-posed [45]. Specific limitations were taken into consideration when applying LDA. A corpus with a few reviews would have provided no theoretical guarantee that good topic identification would occur. If documents are too short, then estimates of word distributions could be biased [43]. In practice, the number of topics should be informed by the application and must be evaluated from the perspective of parsimony [83]. If topics are not concentrated at a compact set of words or reviews as a compact set of topics, then topic models might be misleading [77]. If a single review could be non-uniquely constructed from different combinations of topics, topic models are not expected to show consistent performance.

4. Data

Data was obtained by data scraping two pharmaceutical drug review sites. The data consisted of two data sets on patient reviews on prescription and over-the-counter medications. The data described in Tables 1 and 2 contained three separate reviews of the medication's benefits, side effects and overall experience. Ratings for the medication's benefits and side effects were on a 5-point scale. Ratings for the overall experience were on a 10-point scale. Data from the first data set included ratings of side effects and effectiveness (benefits) but no overall rating. As such, overall rating of the medication was not used as an outcome variable in this research. Instead, the focus was on the ratings of benefits and side effects. The second data set described in Table 2 included a 10-level rating of overall patient satisfaction, the date on which the medication review was submitted, and a indication of the number of other people who found the review helpful.

5. Methods

5.1. Data preprocessing

The raw text of the drug reviews was preprocessed with text mining and natural language processing operations. Preprocessing transformed the raw text from an irregular to a more regular structure. Regularizing the structure of the reviews was crucial for the identification of text

Table 1
Description of data in dataset 1.

Variable	Type	Description
DrugName	Categorical	Name of drug
Condition	Categorical	Name of health condition
BenefitsReview	Text	Review of benefits
SideEffectsReview	Text	Review of side effects
OverallComments	Text	Overall patient comments
SideEffects	Categorical	5-point side effect rating
Effectiveness	Categorical	5-point effectiveness rating

Table 2
Description of data in dataset 2.

Variable	Type	Description
DrugName	Categorical	Name of drug
Condition	Categorical	Name of health condition
BenefitsReview	Text	Review of benefits
SideEffectsReview	Text	Review of side effects
OverallReview	Text	Patient overall review
Rating	Numerical	10-level patient rating
SideEffects	Categorical	5-point side effect rating
Effectiveness	Categorical	5-point effectiveness rating
Date	Date	Date of review entry
UsefulCount	Numerical	Number of users who found review useful

features relevant for and usable by machine learning algorithms used later in this research. Inspection of the preprocessing pipeline in Table 3 highlights the removal of noisy text elements not expected to be useful for the classification algorithms. Two examples that illustrate how raw reviews were regularized by the preprocessing pipeline are given in Table 4.

5.2. Feature extraction and text vectorization

Partially structured text data is most effectively used in mathematical models when it is transformed into a well-structured feature space. Feature extraction is a method to transform the raw text into a numerical representation comprised of a unique combination of word-level features [84]. A bag-of-words (BOW) model was adopted in this research because of its theoretical interpretability and practical efficiency as the means for numerically representing the drug reviews. In the BOW model, each review was represented as a multiset or collection of words from a vocabulary built from the drug reviews. As a syntactic representation model, BOW does not automatically capture word semantics nor word sequence [85]. A feature in BOW model can be generalized to any contiguous sequence of n terms known as an n-grams: unigrams are one-word sequences, bigrams are two-word sequences, and trigrams are three-word sequences [86]. The vocabulary in this research was designed to be the words that survived the preprocessing. Reviews were then mapped to a vector space of dimension equal to the number of n-grams. In vectorized form, a review is simply a vector with length equal to the vocabulary size. An encoding or weighting scheme dictated how the n-gram features were numerically mapped. More than one encoding scheme was utilized to quantify term frequency and inverse document frequency. A document-term matrix (DTM) was formed as the full numerical representation of the drug review corpus, rows of DTM are individual reviews whereas columns are the chosen n-gram. A DTM structure is required as input to many of the algorithms implemented.

6. Exploratory analysis

Characterization of the textual properties of the drug reviews provided a general overview of the collection and supplied an initial assessment of the downstream challenges of the review rating classification problem. Exploratory analyses uncovered properties of n-grams that influenced the nature of the classification. For instance, n-grams were shared between the rating levels of both the reviews for benefits and side effects. Examples of shared unigrams included “pain”, “sleep”, “weight”, and “treatment”. Sharing of unigrams between the rating classes was expected to complicate the classification problem. Examples of bigrams frequently shared between rating classes were “blood pressure”, “weight gain”, “chest pain”, and “stomach pain”. Additional exploration included understanding categories of words with shared grammatical properties, known as parts-of-speech (POS), and that govern aspects of the dependencies between words in the drug reviews [87,88]. The POS system of English includes eight categories: noun, pronoun, adjective, verb, adverb, preposition, conjunction, interjection, and article. Focus was placed on nouns, adjectives, and verbs which delineate entities, characteristics of entities, or actions by entities, respectively [89]. It was found that 93% of the words retained after preprocessing were either nouns, verbs, or adjectives. A major benefit of executing POS tagging in the preprocessing was that n-grams that would have deteriorated the performance of the machine learning algorithms were identified and removed prior to modeling. These uninformative terms included units of time (day, week, month, year, and time), phases of the day (day, night, and afternoon), days of the week, ubiquitous verbs (make and take), and locations (hospital, clinic, house, and home).

6.1. Model evaluation

Classification performance was evaluated with several evaluation metrics. Multiple binary classifications were pooled into aggregated metrics using a binary one-vs-all approach where one rating serves as the positive class and the other ratings serve as the negative [90]. Two aggregations were used, a macro-average, which gives an average over rating classes, and a micro-average, which computes a metric on the sums of prediction results. Macro-average scores were not biased by imbalanced rating classes, since they ignore class frequency and assign equal weight to each class. Micro-averaged scores were biased by class imbalance, since they assign equal weight to each document [91]. As such, the micro-averages were paired with an evaluation metric not sensitive to rating class imbalance. Two macro-aggregated metrics that incorporated the $F1$-score [92] over K classes were utilized,

$$F1_{macro} = \frac{1}{K}\sum_{i=1}^{K} F1\left(TP_i, FP_i, FN_i\right), \tag{1}$$

which was the average of the individual rating per-class $F1$ scores, and

$$F1_{micro} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} TP_i + \frac{1}{2}\left(\sum_{i=1}^{K} FP_i + \sum_{i=1}^{K} FN_i\right)} \tag{2}$$

where TP denotes true positive, FP false positive, TN true negative, and FN false negative classification outcomes. A TP and TN are when the ML algorithm makes a correct classification, whereas FP and FN represented classification errors. In addition to $F1_{macro}$ and $F1_{micro}$, a micro-averaged Matthews correlation coefficient (MCC) [93] was used,

$$MCC_{micro} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{3}$$

where $TP = \sum_{i=1}^{K} TP_i$, $TN = \sum_{i=1}^{K} TN_i$, $FP = \sum_{i=1}^{K} FP_i$, and $FN = \sum_{i=1}^{K} FN_i$ are net classification outcomes, that is, sums over the outcomes after each rating class was treated as the positive class and all others as negative. The MCC_{micro} metric was resistant to imbalance in rating classes by its inclusion of all classification outcomes through MCC [93,94]. A final metric called kappa, derived from Cohen’s kappa statistic, considers the observed agreement with respect to a random guess classifier and is interpreted as a comparison of overall accuracy to the expected accuracy under a random guess [95]

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

Table 3
Data preprocessing pieline to transform drug reviews into more structured representation.

Step	Operation	Example
1	Replace contractions	Don't → Do not
2	Replace word elongations	Whyyyy → Why
3	Replace symbols	@ → at
4	Replace ratings	1/10 → extremely below average
5	Replace ordinals	1st → first
6	Replace numerics	10 → ten
7	Replace kern based spacing	The B O M B → the bomb
8	Replace emoticons	:-) → smiley
9	Replace dates	4/12/2020 → April twelfth two thousand twenty
10	Create corpus	–
11	Lower case	GrEaT → great
12	Remove punctuation	drug, no help → drug no help
13	Remove numbers	3 pills → pills
14	Stemming	suffering → suffer
15	Lemmatization	better → good
16	Remove stop words	a great experience → great experience
17	Custom word removal	doctor, drug, pharmacy, and hospital
18	Spelling correction	adverage → average
19	Strip whitespaces	a ssi t → assist
20	Transform into character vector	–

Table 4
Examples of text before (input) and after (output) the preprocessing pipeline in Table 3.

Example	Input	Output
Benefit	I noticed significant smoothing and softening of lines on face, lightening of uneven coloring, reduction of frequency of blemishes, reduction of oiliness of skin and some tightening of the tissues of the facial skin. I would hope that the effects are long-lasting, but I have only used it for 2 months now.	Notice significant smooth soften line face lighten uneven color reduction frequency blemish reduction oiliness skin tighten tissue facial skin hope long lasting month
Side-Effect	Constant; pulse pounding throughout body, flushed face, odd-looking eyes (dilated),high blood pressure (systolic range 145–194;),fatigue, severe insomnia, nervousness, inability to concentrate, mild vertigo (spill,drop,lurch, fall). Withdrawl symptm (fatigue, tinnitus, leg aches) continue after 4 weeks, though improving gradually increasing over time :panic attacks, anxiety.	Constant pulse pound body flush face odd looking eye dilate high blood pressure systolic range fatigue severe insomnia nervousness inability concentrate mild vertigo spill drop lurch fall withdrawal fatigue tinnitus leg ache continue week improve gradually increase time panic attack anxiety

where p_o is the observed proportionate agreement between a random guess and the trained classifier and p_e is the expected proportionate agreement between the classifiers.

7. Results and discussion

Reviews for the benefits and side effects of the drugs were encoded as document-term matrices (DTM) within the bag-of-words (BOW) model. Feature extraction and representation was done for unigrams and bigrams. Weighting schemes for the DTMs of all supervised learning models, besides the naïve-Bayes (NB) model, included term-frequency-inverse-document-frequency (tf-idf) and term frequency (tf). Reviews with none of the extracted features were dropped from analysis, as their inclusion inflated the sparsity of the DTMs. For the NB model, a Boolean weighting scheme was implemented. Usage of the tf-idf or tf weighting for the NB model generated unstable prediction outcomes dependent on the random initialization of the test and train data. Cosine similarity was used as the distance function in cluster analysis of the review documents because of high dimensionality of the unigram and bigram feature spaces, sparsity of the DTMs, and unequal length of reviews. Latent-Dirichlet allocation was conducted on the tf-idf DTMs.

Results for each class of supervised and unsupervised models will be presented. Results are provided for two versions of the multi-class problem. Version one is a 3-class classification where reviews from the original class set $\{0,1,2,3,4\}$ are mapped to a reduced set $\{0,1,2\}$ where $\{0,1\} \rightarrow \{0\}$, $\{1\} \rightarrow \{1\}$, and $\{3,4\} \rightarrow \{2\}$. Version two is a 5-class classification over the original class set $\{0,1,2,3,4\}$. The difficulty of classification was monotonic in the number of classes. Reviews in the merged classes were determined to have shared content. Positive reviews, those in subclass set $\{3,4\}$, were slightly more numerous at a proportion of 0.71 than the other $\{0,1,2\}$ set, but this was considered minor. Thus, oversampling or undersampling was not used. Instead, the evaluation metrics were designed to be less sensitive to the minor data imbalance. For supervised models, a $70-30$ test-train split was conducted on the raw dataset for both the benefits and side effects reviews. It was decided that a train set with 70% of observations would provide a decent amount of sampling diversity to the models and a 30% test set would provide a balance of observations both seen and unseen. Repeated 10-fold cross validation was implemented during the training phases.

7.1. Unsupervised models

Three clustering models - k-means, agglomerative, and density-based spatial clustering (DBSCAN) - were employed in a stacked architecture, where cluster membership for the k-means algorithm was provided by a hard-vote from the agglomerative and density-based models. The stacked architecture synergized the three models: k-means suffers from the need for a pre-specified number of clusters, agglomerative clustering can handle the problem of a pre-specified number of clusters but is sensitive to outliers, and DBSCAN can handle outliers and abnormally shaped clusters. Two techniques were used to arrive at the optimal number of clusters for data. The elbow method which finds the number of clusters that minimizes the distance of each review from cluster centers suggested three to nine clusters. The silhouette method, which measures how well each review fits within its assigned cluster suggested two-three clusters. All three clustering methods yielded cluster assignments with significant overlapping boundaries. This was evidence that the reviews were not easily separable based on unigram features and that supervised classification would need to be carefully pursued.

Topic modeling with latent Dirichlet allocation (LDA) was conducted to ascertain the topical structure of the drug review corpus. Although no single topic was expected to be unique to a single document and no single word unique to a single topic, a topic model

Table 5
Evaluation metric values for the eight naive-Bayes (NB) models. Macro-averaged F1-score $\left(F1_{macro}\right)$, micro-averaged Matthews correlation coefficient $\left(MCC_{micro}\right)$, and balanced accuracy BA are given for the corresponding value of the Laplace correction (Correction) used when training the naive-Bayes models.

Review	N-gram	Problem	MCC_{micro}	$F1_{macro}$	BA	Correction
Benefits	Unigram	3-class	0.190	0.451	0.592	0.1
Benefits	Unigram	5-class	0.140	0.301	0.566	0.1
Benefits	Bigram	3-class	0.049	0.357	0.528	0.1
Benefits	Bigram	5-class	0.018	0.241	0.501	0.1
Side effects	Unigram	3-class	0.343	0.537	0.654	0.1
Side effects	Unigram	5-class	0.303	0.422	0.639	0.1
Side effects	Bigram	3-class	0.221	0.469	0.605	0.1
Side effects	Bigram	5-class	0.156	0.330	0.582	0.1

revealed coherent topics in the reviews, which informed the classification problem. Three metrics were evoked to determine the best number of topics. First, maximization of the likelihood of a topic model over the number of topics by the Griffith's metric [96]. Second, optimization of the best topic structure as maximized distances between topics [82]. Third, minimization of Kullback–Leibler divergence over topic number [83]. It was found that all metrics would be artificially optimized as the number of topics went to infinity. However, allowing too many topics would have promoted redundancy among topics and an inability to attach meaning to the topics. On the other hand, too few topics yielded LDA models that had low likelihoods, overlapping topic densities, and large divergences. Ten topics were determined to achieve a good balance on all three metrics and admitted reasonably interpretable topics. Two LDA models were then fitted to the benefits and side effects reviews with unigram features. The five unigrams with the greatest probability in a topic were used to define the topic. The topics were "symptom relief", "depression", "skin symptoms", "allergy symptoms", "blood pressure symptoms", "effects on employment", "mood quality", "sleep quality", "magnitude of symptoms", and "pain". These results provided evidence that the semantic structure of the drug review corpus was coherent enough to support rating class discrimination by the supervised models.

7.2. Supervised

The first supervised model investigated was naive-Bayes (NB) which served as a baseline model. It was possible for a drug review to appear in the training data but not in the testing data. In this scenario, naïve-Bayes (NB) would assign zero probability to this review. Laplace smoothing was employed to regularize the NB model, with a small adjustment to each n-gram so that all had non-zero probability. A Laplace correction of 0.1 was determined to provide NB models with good predictive performance as measured by the kappa statistic. A linear grid search was incorporated into 10-fold repeated cross-validation to train an NB models on unigram and bigram features for the 3-class and 5-class problem. Model evaluation metrics for each of these models are provided in Table 5. One additional metric, balanced accuracy (BA), was incorporated as a more commonly understood reference to the uncommon micro-averaged Matthews' correlation coefficient MCC_{micro} and macro-averaged $F1$-score $F1_{macro}$. Balanced accuracy incorporates the accuracy of a model with adjustment for any class imbalances, it gives how much information the model provides over-and-above random guessing by the most common rating class. A few observations were made from the results in Table 5. First, unigrams appeared more informative than bigrams as unigram-based models outperformed their bigram counterparts. This was unexpected and could be due to several factors, such as excessive noise in the bigrams. Second, models based on reviews from side effects demonstrated higher performance than the models derived from the reviews of benefit. This was evidence that the language in the reviews of side effects was more discriminative of the rating of the reviewed drug.

Table 6
Evaluation metric values for the non-linear support vector machine (SVM) with radial-basis function kernel (RBF) models. Macro-averaged F1-score $\left(F1_{macro}\right)$ is given for the maximum micro-averaged Matthews correlation coefficient $\left(MCC_{micro}\right)$ attained. Values for the SVM-RBF hyperparameters of cost C and gamma γ are given.

Review	N-gram	Problem	MCC_{micro}	$F1_{macro}$	C	γ
Benefits	Unigram	3-class	0.206	0.450	10	1e−4
Benefits	Unigram	5-class	0.133	0.309	10	1e−4
Benefits	Bigram	3-class	0.114	0.310	1	0.001
Benefits	Bigram	5-class	0.093	0.249	100	1e−5
Side effects	Unigram	3-class	0.260	0.493	1e4	1e−5
Side effects	Unigram	5-class	0.222	0.375	100	1e−5
Side effects	Bigram	3-class	0.207	0.454	1000	1e−5
Side effects	Bigram	5-class	0.105	0.273	1e4	1e−5

Table 7
Evaluation metric values for the non-linear support vector machine (SVM) with radial-basis function (RBF) kernel models. Micro-averaged Matthews correlation coefficient $\left(MCC_{micro}\right)$ value is given for the maximum macro-averaged F1-score $\left(F1_{macro}\right)$ attained. Values for the corresponding SVM-RBF hyperparameters of cost C and gamma γ are given.

Review	N-gram	Problem	MCC_{micro}	$F1_{macro}$	C	γ
Benefits	Unigram	3-class	0.206	0.450	10	1e−4
Benefits	Unigram	5-class	0.133	0.309	10	1e−4
Benefits	Bigram	3-class	0.102	0.396	1000	1e−5
Benefits	Bigram	5-class	0.093	0.249	100	1e−5
Side effects	Unigram	3-class	0.260	0.493	1e4	1e−5
Side effects	Unigram	5-class	0.222	0.375	100	1e−5
Side effects	Bigram	3-class	0.207	0.454	1000	1e−5
Side effects	Bigram	5-class	0.104	0.276	1000	1e−4

A non-linear support vector machine (SVM) with a radial basis kernel (RBF) was used. Two hyperparameters were involved in the training of the SVM-RBF – gamma (γ) and cost (C). Hyperparameter γ controlled the influence of a single training instance with γ and influence inversely related. Hyperparameter C acted as a regularization parameter for the SVM-RBF models. Hyperparameter tuning of C and γ occurred with a grid search over $C \in \left(0, 1e^{10}\right)$ and $\gamma \in (0, 0.1)$ where SVM-RBF models were fitted to a training set and evaluated according to $F1_{macro}$ and MCC_{micro} on a test set. Training and testing data were generated from random resampling of the original data set. Prediction error varied with γ and C. Smaller values of gamma facilitated models with better generalization error. Extreme values of either C or γ caused a deterioration of generalization error. Results for the SVM-RBF model with the best MCC_{micro} and its corresponding $F1_{macro}$ values are given in Table 6 whereas models with the highest $F1_{macro}$ and corresponding MCC_{micro} are given in Table 7. Similar models were chosen regardless of if MCC_{micro} or $F1_{macro}$ was used to evaluate, evidence that both are doing good jobs at evaluation to models for the reviews of side effects demonstrated higher performance than their counterparts on the reviews of benefits. It was surprising that bigram models displayed worse generalization error when compared to unigram analogues. A possible explanation was that non-optimal feature selection was done, as informative bigrams may have been eliminated during data preprocessing. Non-linear SVMs are highly sensitive to feature and parameter choice, and so further hyperparameter tuning and feature engineering would be recommended in real-world applications.

The capacity to capture the complex regularities between word features and ratings in the reviews is what motivated random forests (RF) with CART decision trees. The relationship between number of trees and classification power of the RF models was found to be unique based on aspects of the problem such as if it were a 3- or 5-class problem and if unigrams or bigrams were used. Much of the change in classification performance occurred between $(0 - 200]$ trees as can be seen in Tables 8 and 9. Prediction performance of the unigram models trained from the benefits reviews showed greater performance than for the reviews of side effects, an inversion of the results for the SVM and NB models. Bigram models underperformed the unigram equivalents,

Table 8
Evaluation metric values for the random forest (RF) models with CART decision trees. Macro-averaged F1-score $(F1_{macro})$ is given for the maximum micro-averaged Matthews correlation coefficient (MCC_{micro}) attained. Values for the RF hyperparameters of number of trees $Ntrees$ are given.

Review	N-gram	Problem	MCC_{micro}	$F1_{macro}$	$Ntrees$
Benefits	Unigram	3-class	0.755	0.822	500
Benefits	Unigram	5-class	0.698	0.781	500
Benefits	Bigram	3-class	0.316	0.570	50
Benefits	Bigram	5-class	0.291	0.418	100
Side effects	Unigram	3-class	0.573	0.707	200
Side effects	Unigram	5-class	0.514	0.597	50
Side effects	Bigram	3-class	0.422	0.637	500
Side effects	Bigram	5-class	0.323	0.467	10

Table 9
Evaluation metric values for the random forest (RF) models with CART decision trees. Micro-averaged Matthew's correlation coefficient (MCC_{micro}) is given for the maximum macro-averaged $F1$-score $(F1_{macro})$ attained. Values for the RF hyperparameters of number of trees $Ntrees$ are given.

Review	N-gram	Problem	MCC_{micro}	$F1_{macro}$	$Ntrees$
Benefits	Unigram	3-class	0.755	0.822	500
Benefits	Unigram	5-class	0.698	0.781	500
Benefits	Bigram	3-class	0.316	0.570	50
Benefits	Bigram	5-class	0.291	0.418	100
Side effects	Unigram	3-class	0.570	0.703	100
Side effects	Unigram	5-class	0.507	0.607	500
Side effects	Bigram	3-class	0.419	0.639	50
Side effects	Bigram	5-class	0.323	0.467	10

an unexpected result (see Tables 8 and 9). It was possible that informative unigrams or bigrams were missed as part of the preprocessing and RF data resampling. The number of features to randomly sample at each node was set to the square root of the number of features, and amounted to subspaces of about 56 features for the unigram models and 70 features for the bigram models.

8. Conclusion

Optimization of medication therapy partly depends on a maximization of benefits and minimization of side effects from medications as perceived by the user. Personalized drug therapies could be formulated from insights learned from patient reported outcomes on benefits and side effects. A patient's evaluative feedback on a drug becomes informative when it can be composed in his/her natural speaking language without being constrained by questions he/she may not understand or words that are outside of his/her dictionary. This work utilized user-written reviews on the benefits and side effects of medications to train and test supervised learning models that could classify a user's review into a 5-point rating scale using only the text as input. Raw data from the reviews was highly irregular and so unsupervised clustering and topic modeling was used with part-of-speech tagging and text mining to establish a robust data preprocessing procedure. The cluster analysis and topic modeling offered the secondary advantage of characterizing the distinguishability of reviews into rating classes, thus giving an early indication of the challenge associated with supervised classification.

Topic models generated through latent Dirichlet allocation revealed that the medication reviews were well modeled as random mixtures of well-defined latent topics. Some of the topics discovered included "sleep quality", "the opportunity to get back to work", and "weight gain". Such topics can be used to design more structured interviews to be given to patients or drug trial participants for more informative feedback on medications. Cluster analysis revealed the existence of different drug review clusters that had overlapping boundaries. The two unsupervised approaches suggested that partitioning by condition or medication prior to supervised analysis could improve classification performance. This follows also from the idea that different health conditions and types of medications are likely to elicit different types of responses with different features such as word choice. The unsupervised learning and parts-of-speech tagging isolated those words most informative to the prediction of satisfaction levels.

Supervised classification with naïve-Bayes, non-linear support vector machine with radial basis function kernels, and random forest algorithms yielded models with moderate out-of-sample prediction accuracy. Random forest outperformed the other two on macro-averaged F1-score and micro-averaged Matthews correlation coefficient. Random forest balanced precision and recall with an overall higher macro-averaged F1-score. Classification on the side effects data resulted in model performance at or above that observed for the data on drug review benefits. Possible explanations of this included that side effects encouraged more complete and informative reviews or word-level features are more distinct for them. Bigram features displayed greater generalization error when compared to unigram features across most of the models. This was surprising, as bigrams have a wider syntactic-level scope and were expected to capture more information. An implication of this is that noise becomes elevated when irrelevant single word features are combined into bigrams. The joint supervised and unsupervised learning gives researchers and practitioners a method to find associations between language features of medication reviews and the predictability of satisfaction with the medication. Furthermore, those benefits or side effects found to be most predictive of satisfaction provide a shared language for patients to provide evaluative feedback essential to fully personalized medicine.

Shortcomings of this work included those identified in earlier sections as well as the need to train, test, and validate the methods on additional data sources. A shortage of data can result in models that underfit or overfit the phenomena under study and yield model performance that is not robust to slight changes in the data. Hyperparameter tuning was also limited in scope by the data limitations. Other data sources represent a fruitful avenue for replication and investigation of more data-hungry models such as deep learning neural networks and large language models. Real-world applications of this research must make sure to use hyperparameter tuning over the new data, and not just use the values in this research, otherwise biased conclusions could ensure. Data was also limited by what was known about the patients who wrote the reviews. It is believed that the data could be augmented with a patient's medical history, social and demographic characteristics, and financial facts. Collection of this data would need to occur alongside efforts to extract reviews from the patients. The reason for this data augmentation is that patients may review drugs differently based on medical, social, demographic, and economic characteristics. If groups of patients defined by their characteristics are suspected to rate differently, then analysis on subsets of the data could isolate more signal and reduce noise from the other patients. A wider sample of drugs and conditions would also allow for subgroup analysis within drug and condition groups, where analysis on distinct groups could isolate more signal. The benefits and side effects reviews were classified separately.

Future research should examine the merits of using automated summarization to generate a single summary over the benefit, side effects, and general reviews, which would then serve as input to text analysis. The extra summarization layer could offer advantages through a stronger noise reduction layer or disadvantages through the loss of informative features. Feedback from patients or medical practitioners may also be given at different times, enabling a human-in-the-loop learning which might serve as a method to adjust for temporal disturbances to the prediction of medication satisfaction. Automated summarization over reviews from more than one user, but for a single medication, could be relevant to understand a population-level satisfaction with a single medication. Such population level understanding would be highly pertinent to pharmaceutical companies, insurance companies, and medical providers with large numbers of patients.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] I. Inayat, S. Salim, S. Marczak, M. Daneva, S. Shamshirband, A systematic literature review on agile requirements engineering practices and challenges, Comput. Hum. Behav. 51 (2015) 915–929.

[2] M. Park, J. Yoo, Benefits of mass customized products: moderating role of product involvement and fashion innovativness, Heliyon 4 (2018).

[3] T. Randall, C. Terwiesch, K. Ulrich, Research note - user design of customized products, Mark. Sci. 26 (2007) 268–280.

[4] D. Whitcomb, What is personalized medicine and what should it replace, Nat. Rev. Gastroenterol. Hepatol. 9 (2012) 418–424.

[5] D. Hayes, H. Markus, R. Leslie, E. Topol, Personalized medicine: risk prediction, targeted therapies and mobile health technology, BMC Med. 12 (2014).

[6] I. Dolgopolov, M. Rykov, The evolution of personalized medicine: a literautre review, Res. Pract. Med. J. 9 (2022).

[7] M. Flores, G. Glusman, K. Brogaard, N. Price, L. Hood, P4 medicinc: how systems medicine will transform the healthcare sector and society, Pers. Med. 10 (2013) 565–576.

[8] M. Sanzo, L. Cipoloni, M. Borro, R.L. Russa, A. Santurro, M. Scopetti, M. Simmaco, P. Frati, Clinical applications of personalized medicine: A new paradigm and challenge, Curr. Pharmaceut. Biotechnol. 18 (2017) 194–203.

[9] W. Alamgir, A. Mohyuddin, Healthcare analytics: Applications and challenges, Life Sci. 3 (2022).

[10] F. Vogenberg, C. Barash, M. Pursel, Personalized medicine: Part 1: Evolution and development into theranostics, P&T 35 (2010) 560–576.

[11] K. Olson, A comprehensie review on healthcare data analytics, J. Biomed. Sustain. Healthc. Appl. 3 (2023) 95–105.

[12] M. Joyner, N. Paneth, Seven questions for personalized medicine, JAMA 314 (2015) 999–10000.

[13] J. Anaya, C. Duarte-Rey, J. Sarmiento-Monroy, D. Bardey, J. Castiblanco, A. Rojas-Villarrada, Personalized medicine. Closing the gap between knowledge and clinical practice, Autoimmun. Rev. 15 (2016) 833–842.

[14] I. Pramanik, R. Lau, A. Azad, S. Hossain, K. Chowdhury, B. Karmaker, Healthcare informatics and analyics in big data, Expert Syst. Appl. 152 (2020).

[15] R. Benjamin, Medication adherence: Helping patients take their medicines as directed, Public Health Rep. 127 (2012) 2–3.

[16] A. Iuga, M. McGuire, Adherence and health care costs, Risk Manage. Healthc. Policy 7 (2014) 35–44.

[17] S. Stewart, Z. Moon, R. Horne, Medication nonadherence: health impact, prevalance, correlates and interventions, Psychol. Health 38 (2022) 726–765.

[18] M. Lemstra, C. nwankwo, Y. Bird, J. Moraaros, Primary nonadherence to chronic disease medications: a meta-analysis, Patient Prefer. Adherence 12 (2017) 721–731.

[19] J. Hugtenburg, L. Timmers, P. Elders, M. Vervloet, L. van Dijk, Definitions, variants, and causes of nonadherence with medication: a challenge for tailored interventions, Patient Prefer. Adherence 7 (2017) 675–682.

[20] S. Hellman, D. Hellman, Of mice but not men - problems of the randomized clinical trial, N. Engl. J. Med. 324 (1991) 1585–1589.

[21] S. Pocock, S. Assman, L. Enos, L. Kasten, Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat. Med. 21 (2002) 2917–2930.

[22] C. Hulsen, Data science in healthcare: COVID-19 and beyond, Int. J. Environ. Res. Public Health 19 (2022).

[23] A. Rahmani, E. Yousefpoor, M. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, R. Naqvi, Machine learning (ML) in medicine: Review, applications, and challenges, Mathematics 9 (2021).

[24] J. Bajwa, U. Munir, A. Nori, B. Williams, Artificial intelligence in healthcare: transforming the pratice of medicine, Future Healthc. 8 (2021).

[25] P. Rajpurkar, E. Chen, O. Banerjee, E. Topol, AI in health and medicine, Nat. Med. 28 (2022) 231–238.

[26] M. Desai, M. Boulos, G. Pomann, G. Steinberg, F. Longo, M. Leonard, T. Montine, A. Blomkalns, R. Harrington, Establishing a data science unit in an academic medical center: An illustrative model, Acad. Med.: J. Assoc. Am. Med. Colleges 97 (2022) 69–75.

[27] C. Voils, R. Hoyle, C. Thorpe, M. Macijewski, W. Yancy, Improving the measurement of self-reported medication nonadherence, Var. Dissent 64 (2011) 250–254.

[28] J. Xu, B. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang, Federated learning for healthcare informatics, J. Healthc. Inform. Res. 5 (2021) 1–19.

[29] J. Rennie, L. Shih, J. Teevan, D. Karger, Tackling the poor assumption of naive Bayes text classifiers, in: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, 2003, pp. 616–623.

[30] M. Ikonomoakis, S. Kotsiantis, V. Tampakas, Text classification using machine learning techniques, WSEAS Trans. Comput. 4 (2005) 966–974.

[31] W. Zhang, T. Toshida, X. Tang, Text classification based on multi-word with support vector machine, Knowl.-Based Syst. 21 (2008) 879–886.

[32] T. Pranckevicius, V. Marcikevicius, Comparison of naive Bayes, random forest, decision tree, support vector mcahines, and logistic regression classifiers for text reviews classification, Balt. J. Mod. Comput. 5 (2017) 221–232.

[33] K. Kowsari, K. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, Information 10 (2019) 150–163.

[34] A. Tripathy, A. Agrawal, S. Rath, Classification of sentimental reviews using machine learning techniques, Procedia Comput. Sci. 57 (2015) 821–829.

[35] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, 2002, pp. 79–86.

[36] K. Dave, S. Lawrence, D. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in: Proceedings of the 12th International Conference on World Wide Web, vol. 12, 2003, pp. 519–528.

[37] F. Graber, S. Kallumadi, H. Malberg, S. Zaunseder, Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning, in: Proceedings of the 2018 International Conference on Digital Health, 2018, pp. 1585–1589.

[38] A. Shiju, Z. He, Classifying drug ratings using user reviews with transformer-based language, IEEE Int. Conf. Healthc. Inform. 10 (2022) 163–169.

[39] L. Augustyniak, P. Szymanski, T. Kajdanowicz, W. Tuligloqicz, Comprehensive study on lexicon-based ensemble classification sentiment analysis, Entropy 18 (2016).

[40] J. Na, W. Kyaing, C. Khoo, S. Foo, Y. Chang, Y. Theng, Sentiment classification of drug reviews using a rule-based linguistic approach, in: The Outreach of Digital Libraries: A Globalized Resource Network. ICADL 2012, vol. 7634, 2012, pp. 189–198.

[41] M. Sokolova, V. Bobicev, Sentiments and opinions in health-related web messages, Proc. Recent Adv. Nat. Lang. Process. (2011) 132–139.

[42] V. Gopalakrishnan, C. Ramaswarmy, Patient opinion mining to analyze drugs satisfaction using supervised learning, J. Appl. Res. Technol. 15 (2017) 311–319.

[43] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. 3 (2003) 993–1022.

[44] D. Zlacky, J. Stas, J. Juhar, A. Cizmar, Categorization with latent Dirichlet allocation, J. Electr. Electron. Eng. 7 (2014) 161–264.

[45] M. Allahyari, M. Assefi, E. Trippe, G. J, K. Kochut, Brief survey of text mining: Classification, clustering and extraction techniques, Proc. KDD (2017) 1–13.

[46] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, KDD Workshop Text Min. 400 (2000) 1525–1526.

[47] Y. Zhao, G. Karypis, Empirical and theoretical comparisons of selected criterion functions for document clustering, Mach. Learn. 55 (2004) 311–331.

[48] A. Dobson, An Introduction To Generlazied Linear Models, third ed., Hall/CRC, 2008.

[49] A. Agresti, Foundations of Linear and Generalized Linear Models, first ed., Wiley, 2015.

[50] H. Zou, T. Hastie, Regularization and variable selection via elastic net, J. R. Stat. Soc. B 67 (2005) 301–320.

[51] S. Indra, L. Wikarsa, R. Turang, Using logistic regression method to classiy tweets into the selected topics, Int. Conf. Adv. Comput. Sci. Inf. Syst. (2016) 385–390.

[52] Y. Chen, P. Liu, C.-P. Teo, Regularised text logistic regression: Key word detection and sentiment classification for online reviews, 2020, ArXiv.

[53] Y. Wang, Z. Huang, Y. Wang, W. Chen, X. Wei, Research on word classification based on logistic regression and machine learning, Int. Conf. Appl. Math. Modell. Intell. Comput. 12756 (2023).

[54] K. Shah, H. Patel, D. Sanghvi, M. Shah, A comparative analysis of logistic regression, random forest and KNN models for text classification, Augment. Hum. Res. 5 (2020).

[55] R. Li, M. Liu, D. Xu, J. Gao, F. Wu, L. Zhu, A review of machine learning algorithms for text classification, Commun. Comput. Inf. Sci. Cyber Secur. 15066 (2022).

[56] J. Wang, X. Li, An improved KNN algorithm for text classfication, Int. Conf. Inf. Netw. Autom. 2 (2010) 436–439.

[57] Y. Zhao, Y. Qian, C. Li, Improved KNN text classification algorithm with MapReduce implementation, Int. Conf. Syst. Inform. 4 (2017) 1417–1422.

[58] R. Wang, Z. Li, J. Cao, T. Chen, L. Wang, Convolutional recurrent neural networks for text classification, Int. Joint Conf. Neural Netw. (2019) 1–6.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all yo need, Adv. Neural Inf. Process. Syst. 5998–6008 (2017).

[60] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pretraining of deep bidirectional transformers for lanaguage understanding., in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguisitics: Human Language Technologies, vol. 1, 2019, pp. 4171–4186.

[61] A. Bhavani, B.S. Kumar, A review of start art of text classification algorithms, Int. Conf. Comput. Methodol. Commun. 5 (2021) 1484–1490.
[62] D. Lewis, Naive (Bayes) at forty: The independence assumption in information retreival, Mach. Learn.: ECML-98 1398 (1998) 4–15.
[63] G. Kaur, E. Oberai, A review article on naive Bayes classifier with various smoothing techniques, Int. J. Comput. Sci. Mob. Comput. 3 (2014) 864–868.
[64] S. Xu, Bayesian naive Bayes classifiers to text classification, J. Inf. Sci. 44 (2016) 3–14.
[65] J. Gareth, D. Witten, T. Hastie, R. Tibshirani, An Introduction To Statistical Learning, first ed., Springer, 2013.
[66] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, 2016.
[67] H. Bhavsar, M. Panchal, A review on support vector machine for data classification, Int. J. Adv. Res. Comput. Eng. Technol. 1 (2012) 31–45.
[68] B. Kumar, O. Vyas, R. Vyas, A comprehensive review on the variants of support vector machines, Modern Phys. Lett. B 33 (2019).
[69] J. Nalepa, M. Kawulok, Selecting training sets for support vector machines: a review, Artif. Intell. Rev. 52 (2019) 897–900.
[70] C. Bishop, Pattern Recognition and Machine Learning, first ed., Springer, 2006.
[71] L. Brieman, Random forests, Mach. Learn. 45 (2002) 5–32.
[72] L. Brieman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[73] T. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Learn. 20 (1998) 832–844.
[74] M. Skuruchina, R. Duin, Bagging, boostring and the random subspace method for linear classifiers, Pattern Anal. Appl. 5 (2002) 121–135.
[75] D. Xu, Y. Tian, A comprehensive survey of clustering algorithms, Ann. Data Sci. 2 (2015) 165–193.
[76] C.A.C. Reddy, Data Clustering: Algorithms and Applications, first ed., Chapman and Hall/CRC, 2013.
[77] I. Vayansky, S. Kumar, A review of topic modeling methods, Inf. Syst. 94 (2020).
[78] R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining, Int. J. Adv. Comput. Sci. Appl. 6 (2015).
[79] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how BERT works, Trans. Assoc. Comput. Linguist. 8 (2020) 842–866.
[80] M. Grootendorst, Bertopic: Neural topic modeling with a class-based TF-IDF procedure, 2022, ArXiv.
[81] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocations (LDA) and topic modelling: models, applications, a survey, Multimedia Tools Appl. 78 (2019) 15169–15211.
[82] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, A density-based method for adaptive LDA model selection, Neurocomputing 7 (2009) 1775–1781.
[83] R. Arun, V. Suresh, C. Madhaven, M. Murthy, On finding the natural number of topics with latent Dirichlet allocation: Some observations, Adv. Knowl. Discov. Data Min. 6118 (2010) 391–402.
[84] R. Dzisevic, D. Sesok, Text classification using different feature extraction aproaches, IEEE Open Conf. Electr. Electron. Inf. Sci. (2019) 1–4.
[85] Y. Zhang, R. Jin, Z. Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. 1 (2010) 42–43.
[86] A. Figueroa, J. Atkinson, Contextual language models for ranking answers to natural language definition questions, Comput. Intell. 28 (2012) 528–548.
[87] S. Kanakaraddi, S. Nandyal, Survey on parts of speech tagger techniques, Int. Conf. Curr. Trends Towards Converg. Technol. (2018) 1–6.
[88] H. Li, H. Mao, J. Wang, Part-of-speech tagging with rule-based data preprocessing and transformer, Electronics 11 (2022) 56–63.
[89] A. Chiche, B. Yitagesu, Part of speech tagging: a systematic review of deep learning and machine learning approaches, J. Big Data 9 (2022).
[90] N. Seliya, T. Khoshgoftaar, J.V. Husle, Aggregating performance metrics for classifier evaluation, IEEE Int. Conf. Inf. Reuse Integr. (2009) 35–40.
[91] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020, ArXiv.
[92] H. Zhu, I. Bayley, M. Green, Metrics for measuring error extents of machine learning classifiers, IEEE Int. Conf. Artif. Intell. Test. (2022) 48–55.
[93] D. Chicco, M. Warrens, G. Jurman, The matthews correlation coefficient (MCC) is more informative than Cohen's kappa and brier score in binary classification assessment, IEEE Access 9 (2021) 78368–78381.
[94] J. Brown, Classifiers and their metrics quantified, Mol. Inform. 37 (2018).
[95] J. Wang, Y. Yang, B. Xia, A simplified Cohen's kappa for use in binary classification data annotation tasks, IEEE Access 7 (2019) 164386–164391.
[96] T. Griffith, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. USA 101 (2004) 5228–5235.