RESEARCH ARTICLE

Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke

Chulho Kim[1,2,3]*, **Vivienne Zhu**[2,3], **Jihad Obeid**[2,3], **Leslie Lenert**[2,3,4]*

[1] Department of Neurology, Hallym University College of Medicine, Chuncheon, Korea, [2] Medical University of South Carolina, Charleston, South Carolina, United States of America, [3] Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina, United States of America, [4] Department of Internal Medicine, Medical University of South Carolina, Charleston, South Carolina, United States of America

* gumdol52@hallym.or.kr (CK); Lenert@musc.edu (LL)

Abstract

Background and purpose

This project assessed performance of natural language processing (NLP) and machine learning (ML) algorithms for classification of brain MRI radiology reports into acute ischemic stroke (AIS) and non-AIS phenotypes.

Materials and methods

All brain MRI reports from a single academic institution over a two year period were randomly divided into 2 groups for ML: training (70%) and testing (30%). Using “quanteda” NLP package, all text data were parsed into tokens to create the data frequency matrix. Ten-fold cross-validation was applied for bias correction of the training set. Labeling for AIS was performed manually, identifying clinical notes. We applied binary logistic regression, naïve Bayesian classification, single decision tree, and support vector machine for the binary classifiers, and we assessed performance of the algorithms by F1-measure. We also assessed how n-grams or term frequency-inverse document frequency weighting affected the performance of the algorithms.

Results

Of all 3,204 brain MRI documents, 432 (14.3%) were labeled as AIS. AIS documents were longer in character length than those of non-AIS (median [interquartile range]; 551 [377–681] vs. 309 [164–396]). Of all ML algorithms, single decision tree had the highest F1-measure (93.2) and accuracy (98.0%). Adding bigrams to the ML model improved F1-mesaure of naïve Bayesian classification, but not in others, and term frequency-inverse document frequency weighting to data frequency matrix did not show any additional performance improvements.

Citation: Kim C, Zhu V, Obeid J, Lenert L (2019) Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS ONE 14(2): e0212778. https://doi.org/10.1371/journal.pone.0212778

Editor: John Shawe-Taylor, University College London, UNITED STATES

Received: August 23, 2018

Accepted: February 8, 2019

Published: February 28, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This study was supported by the NIH National Center for Advancing Translational Sciences (NCATS) through Grant Number UL1 TR001450 and the SmartState Program in South Carolina.

Competing interests: Drs. Kim, Zhu and Obeid have no competing interests. Dr. Lenert is a member of the Board of Directors of the ATCC.

This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Conclusions

Supervised ML based NLP algorithms are useful for automatic classification of brain MRI reports for identification of AIS patients. Single decision tree was the best classifier to identify brain MRI reports with AIS.

Introduction

Stroke is one of the leading causes of death and morbidity worldwide, and a major health problem according to the Global Burden of Disease study [1, 2]. When estimating the burden of a stroke, the incidence, prevalence, and disability-adjusted life-years (DALYs) of the stroke are combined [1, 3]. However, in most studies, the incidence of stroke is not a true national-level figure, but estimated figures that were taken into account in large-scale population-based cohort study results [4, 5]. Alternately, electronic health records can be used to estimate acute stroke incidence [6, 7]. The medical record contains laboratory data, clinical information, and the International Classification of Diseases (ICD) diagnosis codes. Those codes can simply indicate whether a patient has been admitted for a stroke, but often they cannot accurately distinguish whether the patient was hospitalized for acute symptoms of stroke or other problems arising from stroke [8]. However, through various MRI imaging techniques, we can confirm whether the stroke is ischemic or hemorrhagic, and whether it is acute or chronic [9] In addition, MRI reports are rarely coded at a report reading level, and unstructured data such as text reports and imaging data often contains useful information.

One approach for unlocking the information in text descriptions of MRI readings is natural language processing (NLP). NLP has been actively studied in analyses of unstructured text data, which accounts for a large portion of the medical records such as admission notes, nursing records and discharge summaries [10, 11]. NLP tools can be applied in a rule-based fashion to parse out the meaning of texts, although they are employing both supervised and unsupervised machine learning (ML) algorithms [12] Prior stroke research includes feasibility studies of NLP for predicting a future stroke [13], extracting risk factor information [14], and timely screening for urgent thrombolysis [15]. In addition, several reports have used NLP to predict the progression of cancer or to classify breast pathology by analyzing free text radiology reports [16, 17]. However, no NLP study has occurred to identify patients with acute ischemic stroke (AIS) from radiologic reports of brain MRIs. Our aims were to implement ML algorithms that can automatically identify AIS patients based on the free-text in the patients' brain MRI reports. In addition, we compared the performances of different supervised ML algorithms with a harmonized mean of precision and recall in this classification task.

Materials and methods

Participants and MRI sampling

This is a single center retrospective case control study. The study protocol was approved by the Institutional Review Boards and Ethics Committee at Chuncheon Sacred Heart Hospital (IRB No. 2017–114), with a waiver of informed consent. Our hospital stores entire medical records in a clinical data warehouse, which allowed us to screen all brain MRI reports performed between January 1, 2015 and December 31, 2016. We identified MRI reports that included the conventional stroke MRI sequence. Conventional stroke MRI sequences were T2-weighted image, fluid-attenuated inversion recovery, gradient echo image, diffusion weighted image, apparent diffusion coefficient map and non-contrast time-of-flight magnetic resonance (MR) angiography. MRI reports, which also included perfusion or contrast-enhanced MR sequence, were not excluded from the sample. If a patient had a sequence of multiple MRI examinations of the brain, only the first brain MRIs in for each patient was included. During the study period, one neuroradiologist read all brain MRI images. At the time of MRI reading, the neuroradiologist could access information about the chief complaint or reason for referral of the patients to propose an impression of the reading. Additionally, outside imaging or a past imaging were available for the patient, those images were used as a reference for reading the current brain MRI image.

Annotation of MRI reports

The format of the brain MRI reading is depicted in S1 Fig. All the reports were in English. Of these reports, we collected only text data on the radiologists' descriptions and findings of brain MRI reports, and we specifically excluded the texts on the report's conclusions. We consecutively enrolled patients who were admitted to the hospital within 7 days of neurological symptom onset, had consented to participate in a research registry, and were diagnosed with AIS both clinically and radiologically. The registry contains demographic variables, laboratory data, radiologic lesion information, and all the information related to stroke from symptom to post-discharge, such as onset time, emergency department visit time, stroke subtype, type of acute treatment, early neurologic deterioration, and 3-month functional outcome [18]. However, the neuroradiologist could not access the registry which included consensus information about whether the patient had AIS. The gold standard labeling of AIS relied on previous diagnosis of AIS in a prospective AIS registry. In the registry, ischemic stroke was defined as having the relevant lesion on MRI and acute neurological symptoms lasting more than 24 hours [19]. All brain MR images, which were performed in non-AIS subjects and included more than stroke MR sequences, were used as control groups when comparing the text in the findings section of the reading. The control group included patients who underwent brain MRI for a specific disease, such as brain tumor or intracranial hemorrhage, as well as those who underwent MRI as a health check-up or outpatient evaluation for specific symptoms such as headache or dizziness.

NLP algorithm

We used the open source "quanteda" R package, which classifies texts into 2 groups using NLP algorithms (Fig 1) [20]. In brief, full text brain MRI reading sentences were initially parsed into "tokens," with numbers, punctuations, symbols and hyphens in the original text data removed. Then, we used lowercase lettering, stop word removal, and word stemming to normalize those data [21]. Finally, we obtained the document-feature matrix (dfm), which is a vector representation of tokens that are truncated from the whole text. We used 4 types of dfm vectorization: unigram, unigram + term frequency-inverse document frequency (tf-idf), adding bigram, and adding bigram + tf-idf. Term frequency (tf) is the number of times that a particular word occurs in a document, and document frequency is the count of documents containing a particular word [22]. Inverse document frequency (idf) is the reciprocal of document frequency. For example, idf value is small for common words such as "the", and large for those that are not common. Tf-idf is a way of giving weight to a word vector by multiplying tf by idf. Bigram is a two-word vector that is arranged in a sequential manner, which helps to differentiate a document by the word quantity as well as the word order [23].

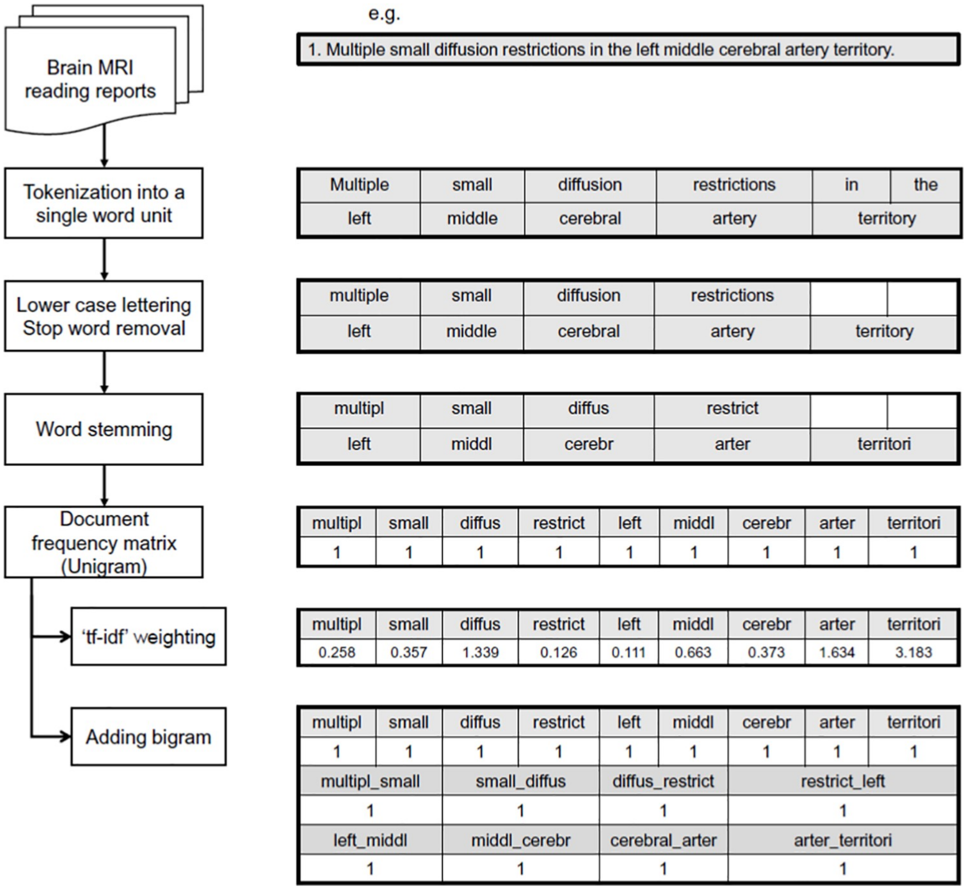
e.g.
1. Multiple small diffusion restrictions in the left middle cerebral artery territory.
Brain MRI reading reports
Tokenization into a single word unit
Multiple | small | diffusion | restrictions | in | the
left | middle | cerebral | artery | territory
Lower case lettering Stop word removal
multiple | small | diffusion | restrictions
left | middle | cerebral | artery | territory
Word stemming
multipl | small | diffus | restrict
left | middl | cerebr | arter | territori
Document frequency matrix (Unigram)
multipl | small | diffus | restrict | left | middl | cerebr | arter | territori
1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1
'tf-idf' weighting
multipl | small | diffus | restrict | left | middl | cerebr | arter | territori
0.258 | 0.357 | 1.339 | 0.126 | 0.111 | 0.663 | 0.373 | 1.634 | 3.183
Adding bigram
multipl | small | diffus | restrict | left | middl | cerebr | arter | territori
1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1
multipl_small | small_diffus | diffus_restrict | restrict_left
1 | 1 | 1 | 1
left_middl | middl_cerebr | cerebral_arter | arter_territori
1 | 1 | 1 | 1

Fig 1. Preprocessing flow chart of "quanteda" natural language processing package.

https://doi.org/10.1371/journal.pone.0212778.g001

Statistical analysis

We performed descriptive analyses of differences between AIS and non-AIS reports. Character lengths of the reports were compared using a Mann-Whitney U test. We used the "keyness plot" to determine which words were frequently used in AIS readings and which words were frequently used in non-AIS. The chi-square value of the plot indicates the frequency of the words appearing in the document, and that value becomes smaller and approaches zero when the words appearing simultaneously in two documents of AIS and non-AIS patients [24].

To classify the two reference standards of AIS and non-AIS, four types of dfm matrix were applied to 4 ML algorithms—binary logistic regression (BLR), naïve Bayesian classification (NBC), single decision tree (SDT) and support vector machine (SVM). We split the text data into training and testing datasets with a ratio of 7:3 and used 10-fold cross-validation to train the models on the training set. We compared the performance of the four algorithms with F1-measure (harmonized mean and precision and recall) and receiver operating characteristic (ROC) curve analysis in classifying AIS and non-AIS reports. The e1071, rpart, and quanteda packages were used to perform all our statistical analyses and ML algorithms; all statistical computing was performed with R (version 3.4.3) [25, 26].

In addition, we performed a qualitative analysis of MRI readings that were misclassified by the best performing ML model. In the case of supervised ML classifiers, it may be important to correct the class imbalance during the training process to reduce the bias and to obtain better performance [27]. Therefore, ML training was performed by random sampling of training data corresponding to each class balanced to 50:50 by setting with a case number (303 vs 303), a control number (1815 vs 1815), or a desired number (5000 in total) [28].

Results

Of all 8,793 brain MRI readings, 4,238 MRIs included more than conventional stroke MRI sequences. A total of 3,024 MRIs was included in the final analysis, excluding those taken more than twice during the study period. Raw data can be downloaded in the Supporting Information File (S1 File). The mean age of the participants and proportion of female were 60.0 ± 17.6 years and 51.7% (1,563 out of 3,024), respectively. During the study period, there were 469 AIS patients were enrolled in the registry; we excluded 37 subjects with an AIS because they did not have enough stroke MRI sequence images, or they only had MRI images from outside the hospital. The test and training data sets included 432 (14.3%) patients with MRI readings that confirmed AIS. The resulting training dataset had 303 AIS and 1,815 non-AIS reports, and the test dataset had 129 AIS and 777 non-AIS reports.

Fig 2 depicts the difference of the text character lengths between AIS and non-AIS reports. MRI reports of AIS patients had a larger amount of text characters versus reports of non-AIS patients (median [interquartile range]; 551 [377–681] vs. 309 [164–396]). We show the 15 most frequently occurring words in the AIS reading and those words in the non-AIS readings, and we summarize them in Fig 3. For example, the word “acute” was used most frequently in AIS reports, followed by “restrictions” and “cortex”. On the other hand, the words “gross”, “abnormal”, and “finding”, which usually represent normal conditions (e.g., “No gross abnormal findings was observed.”), appeared frequently in non-AIS reports.

MRI reading classification by NLP

Of 2,118 randomly selected reports in the training dataset, text preprocessing of MRI reports yielded 1,146 keyword features after removing numbers, punctuations, symbols, hyphens and stop words. When we extracted the keywords using bigram as well as unigram in text

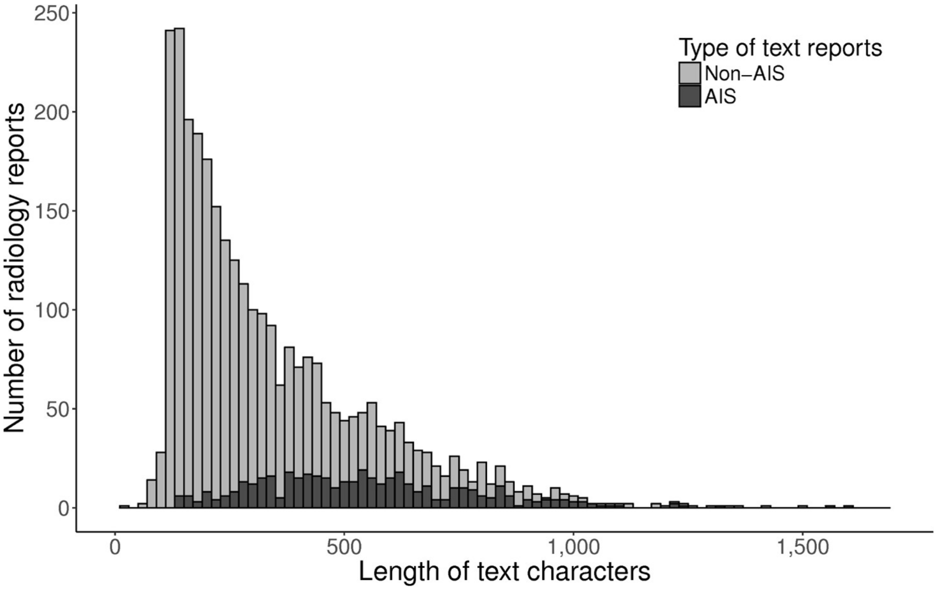

Fig 2. Difference of the text character lengths between AIS and non-AIS reports. AIS, acute ischemic stroke.

https://doi.org/10.1371/journal.pone.0212778.g002

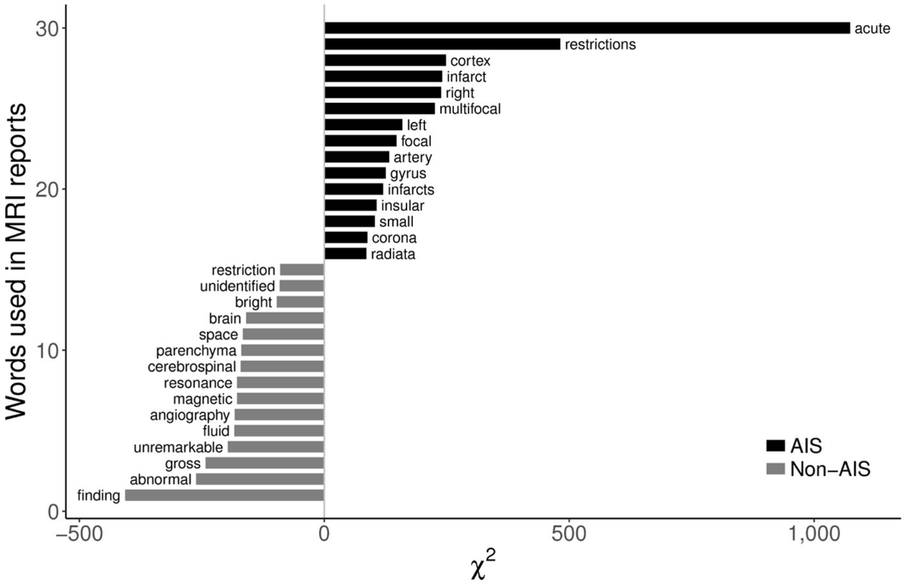

Fig 3. Result of keyness plot analysis of AIS and non-AIS reports. AIS, acute ischemic stroke.

https://doi.org/10.1371/journal.pone.0212778.g003

classification, 9,402 features were obtained and entered into the training dataset and used to predict each ML algorithm. Precision, also known as positive predictive value, was defined as the ratio of true positive over true positive plus false positive, while recall, also known as sensitivity, was defined as the ratio of true positive results in the test over the true positive plus false negative. We presented the performance of each algorithms as the F1-measure (harmonized mean of precision and recall):

$$\text{F1 measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Fig 4 shows a comparison of these performance of each algorithm and detailed results are presented in Table 1. Of all the ML algorithms, the F1-measure of SDT was the highest in unigram classification even if we added bigram or tf-idf weights in the ML model. Adding the bigram to the ML model improved performance in NBC, but not in other models. S2 Fig also shows the area under the ROC of each ML algorithm. Adding the bigram to the ML model, which requires more computational efforts in performing the ML task, could improve the recall slightly, but overall performance of the BLR or SVM was not improved.

Decision tree and error analysis

Performance of SDT produced 93.2 of F1-measure as well as a good accuracy (98.0% in Table 1). The”acut” feature was located in the root node, while the “intracerebr” and “intraventricular” features, which usually imply an intracranial hemorrhage, were located in the internal nodes to distinguish AIS from non-AIS. There were 12 false positive and 6 false negative results for this algorithm, and the relevant explanations for the misclassification are summarized in Table 2.

Model considering class imbalance of training

The training dataset of the single decision tree was composed of 303 AIS cases and 1,815 controls. We used three methods to resolve the class imbalance in decision tree training: over

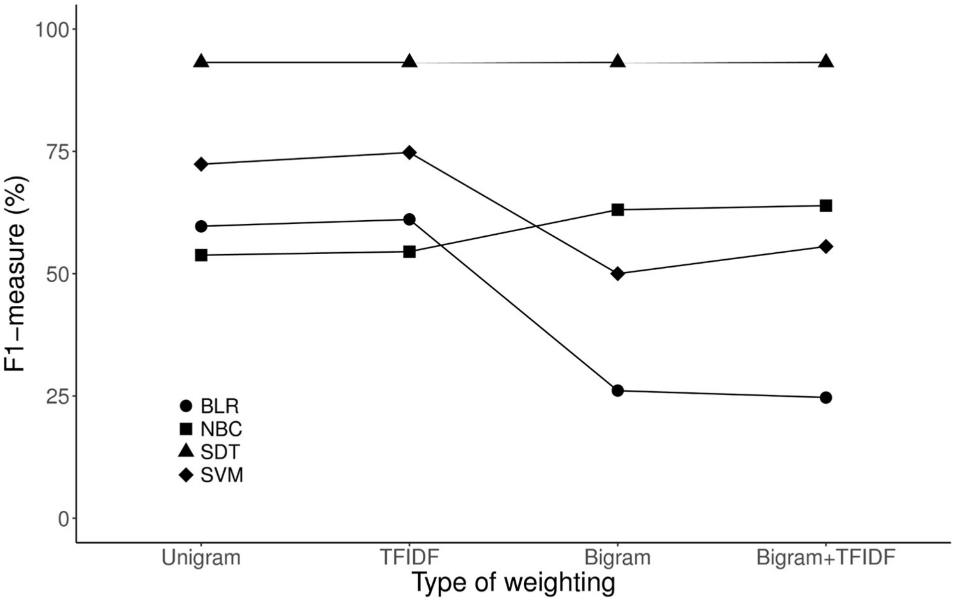

Fig 4. Comparison of ML and NLP algorithms for classifying the brain MRI reports. ML, machine learning; NLP, natural language processing, BLR, binary logistic regression; NBC, naïve Bayesian classification; SDT, single decision tree; SVM, support vector machine; TFIDF, term frequency-inverse document frequency.

https://doi.org/10.1371/journal.pone.0212778.g004

sampling (1,815 vs 1,815), under sampling (303 vs. 303), and fixed number (n = 5,000) sampling (S1 Table). There was no significant change of performance in precision, recall, accuracy, and the F1-measure when we obtained test results after training with those balanced data sets.

Table 1. Results of performance of each machine learning algorithms.

	TP	FP	FN	TN	Total	Sensitivity (Recall)	Specificity	PPV (Precision)	NPV	Accuracy	F1-measure	P for χ^2
BLR unigram	100	106	29	671	906	77.5	86.4	48.5	95.9	85.1	59.7	<0.001
BLR tf-idf	102	103	27	674	906	79.1	86.7	49.8	96.1	85.7	61.1	<0.001
BLR adding bigram	64	298	65	479	906	49.6	61.6	17.7	88.1	59.9	26.1	0.020
BLR adding bigram+tf-idf**	60	297	69	480	906	46.5	61.8	16.8	87.4	59.6	24.7	0.082
NBC unigram	110	170	19	607	906	85.3	78.1	39.3	97.0	79.1	53.8	<0.001
NBC tf-idf	112	170	17	607	906	86.8	78.1	39.7	97.3	79.4	54.5	<0.001
NBC adding bigram	111	112	18	665	906	86.0	85.6	49.8	97.4	85.7	63.1	<0.001
NBC adding bigram+tf-idf	116	118	13	659	906	89.9	84.8	49.6	98.1	85.5	63.9	<0.001
SDT unigram*	123	12	6	765	906	95.3	98.5	91.1	99.2	98.0	93.2	<0.001
SDT tf-idf*	123	12	6	765	906	95.3	98.5	91.1	99.2	98.0	93.2	<0.001
SDT adding bigram*	123	12	6	765	906	95.3	98.5	91.1	99.2	98.0	93.2	<0.001
SDT adding bigram+tf-idf*	123	12	6	765	906	95.3	98.5	91.1	99.2	98.0	93.2	<0.001
SVM unigram	76	5	53	772	906	58.9	99.4	93.8	93.6	93.6	72.4	<0.001
SVM tf-idf	80	5	49	772	906	62.0	99.4	94.1	94.0	94.0	74.8	<0.001
SVM adding bigram	43	0	86	777	906	33.3	100.0	100.0	90.0	90.5	50.0	<0.001
SVM adding bigram+tf-idf	50	1	79	776	906	38.8	99.9	98.0	90.8	91.2	55.6	<0.001

TP, true positive; FP, false positive; FN, false negative; TN, true negative; PPV, positive predictive value, NPV, negative predictive value; BLR, binary logistic regression; tf-idf, term frequency-inverse document frequency; NBC, Naïve Bayesian classification; SDT, single decision tree; SVM, support vector machine.

* the best classifiers.

** the worst classifier.

https://doi.org/10.1371/journal.pone.0212778.t001

Table 2. Error analysis of result of single decision tree in classifying AIS and non-AIS.

Reason for misclassification	FN	FP
Various disease condition could be accompanied with MR diffusion restrictions	3	0
Reading including the recent or old cerebral hemorrhages	3	4
Lesions with diffusion restrictions in MRI but no relevant clinical symptoms	5	0
Miscellaneous	1	2
Total	12	6

AIS, acute ischemic stroke; FN, false negative; FP, false positive.

https://doi.org/10.1371/journal.pone.0212778.t002

Discussion

In our study, NLP algorithms were a useful tool to identify patients with the phenotype of AIS, using unstructured radiologic reports of brain MRIs. Interestingly, SDT-based binary classification showed high precision (91.1%) and recall (95.3%), and additional weighting method for dfm did not show further improvement of several ML algorithms. Error analysis of SDT showed that most of the errors were not caused by NLP or ML algorithms but by the MRI imaging characteristics of the AIS itself. In terms of classification imbalance during the SDT training process, there were no significant differences of the F1-measures of ML predictions when we performed training processes using several class-balanced data.

Since the 1980s when CT equipment in conjunction with X-rays began to be used for the diagnosis of human illness, the development of diagnostic equipment has evolved rapidly. Various imaging techniques have been used to diagnose specific brain diseases, and brain MRI has become an essential tool for the diagnosis of various diseases including AIS [29]. Because MRI images are proliferating at a rapid rate and the MRI reading is an unstructured text data, it is becoming increasingly difficult to classify those diagnostic images manually within a fixed time period. Moreover, it may be inaccurate to classify CNS diseases using diagnostic codes such as the ICD [30, 31], which are usually coded manually. In the case of AIS caused by other main diseases, such as cardiogenic AIS caused by acute myocardial infarction, the stroke diagnosis code may be secondary to the ICD codes. In addition, two studies that analyzed a trend of intravenous thrombolysis after acute ischemic stroke with the ICD-9 codes reported that the ICD-9 codes tended to underestimate intravenous thrombolysis [32, 33]. Therefore, diagnostic codes such as the ICD-9 may return inaccurate search results for certain diseases such as AIS. However, our study demonstrated that information related to an AIS diagnosis could be successfully extracted in large numbers of brain MRI radiology reports using open source NLP and ML algorithms. We suggest that these automated supervised ML and NLP algorithms could be beneficial in classifying a vast amount of brain MRI reports automatically and accurately.

Our NLP-based ML technique makes it possible to classify and extract useful information efficiently in a short period of time from a large amount of text reports. Wright et al. used lexicon-based ML classification for extracting diabetes-related information from 2000 clinical progress notes and reported that SVM using a bag-of-words approach was effective in classifying them as 0.96 of AUROC and 0.93 of the F1-score [34]. Hassanpour et al. suggested that simple structured texts could be sufficiently classified with a bag-of-words model and complex structured texts with lexicon-based information retrieval methods [35]. In our analysis, we applied bag-of-words NLP algorithms to identify AIS reports from a large amount of brain MRI radiology reports, and their algorithmic performances were comparable to other study results [34,36,37]. Our result suggest that the brain MRI radiology report is not a complex structured text. However, further study is needed to determine whether the bag-of-words model is more important than the higher order classification system for multi-class classification.

Usually adding bigram features on a bag-of-words unigram model improves the classification performance because the text itself is the sum of the sequential vectors [38]. However, combined unigram-bigram features did not improve classification performance in our analysis. The reasons why this phenomenon occurred are as follows. First, applying bigram to input vectors produces a large amount of input data. In our model, input vector size increased from 1,146 to 9,402 features. Moreover, performance of the ML classifier depends on the trade-off between false positives and false negatives. Therefore, the large number of word vectors created by adding bigram features to NLP may have contributed to a further reduction in performance in binary classification. Grundmeier et al. suggested that removal words with low frequency in each text from a large number of input features could successfully identify long bone fractures in radiology reports [39]. Second, in the SDT structure, the more important predictors are located near the root node [40]. Grundmeier et al. studied the NLP classification adding bigram features to the random forest classifier, which is an ensemble of decision trees. And they showed that unigram features had higher Gini importance values when compared to bigram features [6]. Therefore, we speculate that the performance of SDT did not improve by adding bigram because unigram features were located in the uppermost node in the decision tree.

Fig 3 shows the results of a keyness plot indicating "keyword" features and comparing their differential associations with an AIS versus a non-AIS group. That representative example illustrates that "keyword searching" can extract information but in an inefficient way when compared to the NLP method. A large number of words expressing stroke lesion were identified in the AIS reports, while those that described normal reading, such as "unidentified bright object", "unremarkable" or "no gross abnormal finding" were located in those of non-AIS. However, the words "restriction" or "restrictions" appeared in both AIS and non-AIS reports. Because word stemming as well as lowercase lettering used in NLP can condense various types of words into a single etymology, it is possible to process text features more efficiently with NLP versus keyword searching in text classification. Doan et al. reported that an NLP tool had a higher sensitivity (93.6% vs. 41.0%) in identifying Kawasaki disease in emergency department notes when compared to a simple keyword research, which suggested that the NLP tool could be a good decision support system for the proper diagnosis in an emergent clinical setting when compared to knowledge-based clinical decision-making alone [41]. Thus, we also showed that text mining using NLP had a high accuracy and efficiency compared to keyword searching.

We found that radiology reports of AIS had a longer length than those of non-AIS. Text length could be an important marker in differentiating ham and spam in supervised text message classification [42]. Several structured data such as age and sex are not included in protected health information identifiers and are readily available from the electronic health record, those structured data contain valuable information related to the risk of developing a particular disease. Therefore, it is expected that additional modeling with unstructured data and selected structured data may have a beneficial effect on the performance of ML algorithms in classifying radiology reports. However, we only used the deidentified unstructured text data for this study; further research is needed to determine the effects of additionally using structured data to assess classification performance.

In our result, we showed that SDT had a higher performance for binary classification than the other ML algorithms. Generally, a decision tree performs well when dealing with discrete or categorical features, while SVM performs well with continuous features [43]. Chen et al. analyzed the performance of an ML algorithm to categorize oncologic response using abdominal CT and MRI reports; those researchers showed that SVM had a higher performance (accuracy = 90.6, F score = 0.81) versus analyses with Bayes point machine, logistic regression, random forest, or neural network [16]. However, the performance of SVM decreased when more than 2,500 features were included in the ML algorithms. We also identified that F1-measure was lower when SVM was performed using an n-gram, which requires more additional features during training, as compared to unigram bag-of-words training.

Also, the performance of SVM is reported to be better than decision tree when classification is performed using imaging data or voice data [44]. Yadav et al. reported that the decision tree showed high performance when binary classification was performed for traumatic brain injury using brain CT readings [45]. Likewise, we found that, to achieve good performance, it may be better to choose decision tree as a classifier if the researchers choose to perform a binary classification using brain or CT or MRI radiology reports. However, the factors affecting the performance of the classifier include the amount of training data, characteristics of those data, and class imbalance, and the type of classifier [43]. Therefore, we should carefully consider characteristics of the data when we select for the ML classifier of NLP algorithms.

The resulting error analysis for SDT was due to the radiological characteristics of disease in the CNS rather than errors in NLP or ML algorithms. Diffusion-restrictive lesion is not only a main MRI characteristic of the AIS lesion, it is also accompanied by hypoxia, excitotoxicity, and perihematomal ischemia of the brain [46]. Other NLP tools such as continuous skip-gram of word2vec [47] and GloVe [48] could take into account order and proximity of the words. It is worth investigating whether these NLP methodologies can reduce the errors seen in our results.

There are several limitations to our study. First, our text corpus was created at a single institution, and therefore, it is not possible to generalize our findings. However, generalizable results could occur if we use those NLP and ML tools for inter-institutional validation in a future study. Second, we only included brain MRI reports with conventional stroke MRI sequence. In clinical practice, full conventional brain MRI sequence could vary depending on the degree of emergency in a given situation, the patient's condition, and the laboratory results. In other words, diffusion only MRI instead of the full stroke MRI sequences would be performed in cases of emergency or when a patient is unstable. Although a diffusion only MRI report is sometimes used to diagnose AIS, that technique does not have all the text features of AIS because the report only includes the description of the diffusion MRI. Therefore, it is important to investigate the characteristics of each institutional radiology report before application of NLP and ML algorithms. Lastly, the performance of ML classifiers could be affected by the class proportions in the training dataset [49]. The proportion of brain MRI reporting in AIS may vary significantly depending on the characteristics of each hospital. However, we obtained results using a balanced dataset, so we can expect that differences in class proportion in the training dataset will not affect the outcome.

Conclusions

Supervised ML and NLP algorithms can successfully classify brain MRI reports for identification of AIS patients. Moreover, these techniques are rapidly developing fields that can automatically classify a vast amount of medical images using deep learning algorithms. However, labeling for the image data is also a challenging problem in the field of image classification. Therefore, the NLP algorithms can be used to label image data for deep learning.

Supporting information

S1 Fig. Format of the brain MRI readings.
(DOCX)

S1 File. Raw data.
(CSV)

S2 Fig. ROC curve analysis for ML classifier according to NLP weighting methods.
(DOCX)

S1 Table. Results of single decision tree for binary classification considering random sampling of training dataset for reducing class imbalance.
(DOCX)

Author Contributions

Conceptualization: Jihad Obeid, Leslie Lenert.

Data curation: Vivienne Zhu, Jihad Obeid.

Formal analysis: Chulho Kim.

Funding acquisition: Leslie Lenert.

Investigation: Chulho Kim.

Methodology: Chulho Kim, Vivienne Zhu, Jihad Obeid, Leslie Lenert.

Project administration: Vivienne Zhu, Jihad Obeid, Leslie Lenert.

Supervision: Jihad Obeid, Leslie Lenert.

Validation: Chulho Kim, Vivienne Zhu.

Writing – original draft: Chulho Kim.

Writing – review & editing: Vivienne Zhu, Jihad Obeid, Leslie Lenert.

References

1. GBD 2015 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the global burden of disease study 2015. Lancet. 2016; 388:1603–1658. https://doi.org/10.1016/S0140-6736(16)31460-X PMID: 27733283

2. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. Lancet. 2016; 388:1459–1544. https://doi.org/10.1016/S0140-6736(16)31012-1 PMID: 27733281

3. Murray CJ, Ezzati M, Flaxman AD, Lim S, Lozano R, Michaud C, et al. GBD 2010: a multi-investigator collaboration for global comparative descriptive epidemiology. Lancet. 2012; 380:2055–2058.

4. Krishnamurthi RV, Barker-Collo S, Parag V, Parmar P, Witt E, Jones A, et al. Stroke incidence by major pathological type and ischemic subtypes in the Auckland regional community stroke studies: changes between 2002 and 2011. Stroke. 2018; 49:3–10. https://doi.org/10.1161/STROKEAHA.117.019358 PMID: 29212738

5. Koton S, Schneider AL, Rosamond WD, Shahar E, Sang Y, Gottesman RF, et al. Stroke incidence and mortality trends in US communities, 1987 to 2011. JAMA. 2014; 312:259–268. https://doi.org/10.1001/jama.2014.7692 PMID: 25027141

6. Willers C, Lekander I, Ekstrand E, Lilja M, Pessah-Rasmussen H, Sunnerhagen KS, et al. Sex as predictor for achieved health outcomes and received care in ischemic stroke and intracerebral hemorrhage: a register-based study. Biol Sex Differ. 2018; 9:11. https://doi.org/10.1186/s13293-018-0170-1 PMID: 29514685

7. Dhamoon MS, Liang JW, Zhou L, Stamplecoski M, Kapral MK, Shah BR. Sex differences in outcomes after stroke in patients with diabetes in Ontario, Canada. J Stroke Cerebrovasc Dis. 2018; 27:210–220. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.08.028 PMID: 28918090

8. Baldereschi M, Balzi D, Di Fabrizio V, De Vito L, Ricci R, D'Onofrio P, et al. Administrative data underestimate acute ischemic stroke events and thrombolysis treatments: data from a multicenter validation survey in Italy. PLoS One. 2018; 13:e0193776. https://doi.org/10.1371/journal.pone.0193776 PMID: 29534079

9. Vilela P, Rowley HA. Brain ischemia: CT and MRI techniques in acute ischemic stroke. Eur J Radiol. 2017; 96:162–172. https://doi.org/10.1016/j.ejrad.2017.08.014 PMID: 29054448

10. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. 2017; 73:14–29. https://doi.org/10.1016/j.jbi.2017.07.012 PMID: 28729030

11. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics. 2016; 36:176–191. https://doi.org/10.1148/rg.2016150080 PMID: 26761536

12. Lacson R, Khorasani R. Practical examples of natural language processing in radiology. J Am Coll Radiol. 2011; 8:872–874. https://doi.org/10.1016/j.jacr.2011.09.010 PMID: 22137006

13. Hung CY, Chen YC, Lai PT, Lin CH, Lee CC. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. Conf Proc IEEE Eng Med Biol Soc. 2017; 2017:3110–3113. https://doi.org/10.1109/EMBC.2017.8037515 PMID: 29060556

14. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics. 2016; 7:26. https://doi.org/10.1186/s13326-016-0065-1 PMID: 27175226

15. Sung SF, Chen K, Wu DP, Hung LC, Su YH, Hu YH. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. Int J Med Inform. 2018; 112:149–157. https://doi.org/10.1016/j.ijmedinf.2018.02.005 PMID: 29500013

16. Chen PH, Zafar H, Galperin-Aizenberg M, Cook T. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging. 2018; 31:178–184. https://doi.org/10.1007/s10278-017-0027-x PMID: 29079959

17. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat. 2017; 161:203–211. https://doi.org/10.1007/s10549-016-4035-1 PMID: 27826755

18. Kim BJ, Park JM, Kang K, Lee SJ, Ko Y, Kim JG, et al. Case characteristics, hyperacute treatment, and outcome information from the clinical research center for stroke-fifth division registry in South Korea. J Stroke. 2015; 17:38. https://doi.org/10.5853/jos.2015.17.1.38 PMID: 25692106

19. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJ, Culebras A, et al. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013; 44:2064–2089. https://doi.org/10.1161/STR.0b013e318296aeca PMID: 23652265

20. Benoit K, Nulty PP. Quanteda: Quantitative analysis of textual data. R package version 0.9. 6–9. 2016;8. Available from: https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

21. Porter MF. An algorithm for suffix stripping. Program. 1980; 14:130–137.

22. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage. 1988; 24:513–523.

23. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. Comput Linguist. 1992; 18:467–479.

24. Culpeper J. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. Int J Corp Linguist. 2009; 14:29–59.

25. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. E1071: Misc functions of the department of statistics (e1071), TU Wien, 2014. R package version 2015:1.6–4. Available from: https://cran.r-project.org/web/packages/e1071/e1071.pdf

26. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the rpart routines. Rochester: Mayo Foundation 2000. Available from: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

27. Kaur P, Gosain A. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. 1st ed. Springer: ICT Based Innov; 2018. pp. 23–30.

28. Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. R Journal. 2014; 6. Available from: https://journal.r-project.org/archive/2014/RJ-2014-008/RJ-2014-008.pdf

29. Fragata I, Alves M, Papoila AL, Nunes AP, Ferreira P, Canto-Moreira N, et al. Early prediction of delayed ischemia and functional outcome in acute subarachnoid hemorrhage: role of diffusion tensor

imaging. Stroke. 2017; 48:2091–2097. https://doi.org/10.1161/STROKEAHA.117.016811 PMID: 28667021

30. Piriyawat P, Smajsova M, Smith MA, Pallegar S, Al-Wabil A, Garcia NM, et al. Comparison of active and passive surveillance for cerebrovascular disease: The Brain Attack Surveillance in Corpus Christi (BASIC) Project. Am J Epidemiol. 2002; 156:1062–1069. PMID: 12446264

31. Baldereschi M, Balzi D, Di Fabrizio V, De Vito L, Ricci R, D'Onofrio P, et al. Administrative data underestimate acute ischemic stroke events and thrombolysis treatments: data from a multicenter validation survey in Italy. PLoS One. 2018; 13:e0193776. https://doi.org/10.1371/journal.pone.0193776 PMID: 29534079

32. Kleindorfer D, Lindsell CJ, Brass L, Koroshetz W, Broderick JP. National US estimates of recombinant tissue plasminogen activator use: ICD-9 codes substantially underestimate. Stroke. 2008; 39:924–928. https://doi.org/10.1161/STROKEAHA.107.490375 PMID: 18239184

33. Adeoye O, Hornung R, Khatri P, Kleindorfer D. Recombinant tissue-type plasminogen activator use for ischemic stroke in the United States: a doubling of treatment rates over the course of 5 years. Stroke. 2011; 42:1952–1955. https://doi.org/10.1161/STROKEAHA.110.612358 PMID: 21636813

34. Wright A, McCoy AB, Henkin S, Kale A, Sittig DF. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. J Am Med Inform Assoc. 2013; 20:887–890. https://doi.org/10.1136/amiajnl-2012-001576 PMID: 23543111

35. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med. 2016; 66:29–39. https://doi.org/10.1016/j.artmed.2015.09.007 PMID: 26481140

36. Zhou Y, Amundson PK, Yu F, Kessler MM, Benzinger TL, Wippold FJ. Automated classification of radiology reports to facilitate retrospective study in radiology. J Digit Imaging. 2014; 27:730–736. https://doi.org/10.1007/s10278-014-9708-x PMID: 24874407

37. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc. 2014; 22:155–165. https://doi.org/10.1136/amiajnl-2014-002768 PMID: 25332356

38. Tan CM, Wang YF, Lee CD The use of bigrams to enhance text categorization. Inform Process Manag 2002; 38:529–546.

39. Grundmeier RW, Masino AJ, Casper TC, Dean JM, Bell J, Enriquez R, et al. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform. 2016; 7:1051–1068. https://doi.org/10.4338/ACI-2016-08-RA-0129 PMID: 27826610

40. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27:130–135. https://doi.org/10.11919/j.issn.1002-0829.215044 PMID: 26120265

41. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, et al. Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes. Acad Emerg Med. 2016; 23:628–636. https://doi.org/10.1111/acem.12925 PMID: 26826020

42. Liu W, Wang T. Index-based online text classification for sms spam filtering. J Comput. 2010; 5:844–851.

43. Ilias Maglogiannis KK, Manolis Wallace, John Soldatos. Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies IOS Press; 2007.

44. Lahmiri S, Dawson DA, Shmuel A. Performance of machine learning methods in diagnosing parkinson's disease based on dysphonia measures. Biomed Eng Lett. 2018; 8:29–39. https://doi.org/10.1007/s13534-017-0051-2 PMID: 30603188

45. Yadav K, Sarioglu E, Choi HA, Cartwright WB 4th, Hinds PS, Chamberlain JM. Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. Acad Emerg Med. 2016 23:171–178. https://doi.org/10.1111/acem.12859 PMID: 26766600

46. Schaefer PW, Grant PE, Gonzalez RG. Diffusion-weighted MR imaging of the brain. Radiology. 2000; 217:331–345. https://doi.org/10.1148/radiology.217.2.r00nv24331 PMID: 11058626

47. Lilleberg J, Zhu Y, Zhang Y. Support vector machines and word2vec for text classification with semantic features. IEEE Cogn Inform Cogn Comput. 2015:136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

48. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. Proc Empir Methods Nat Lang Process. 2014:1532–1543. https://doi.org/10.3115/v1/D14-1162

49. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16:321–357.