Unstructured Medical Text Classification Using Linguistic Analysis: A Supervised Deep Learning Approach

Ahmad Al-Doulat[1], Islam Obaidat[2], and Minwoo Lee[3]

[1,2,3]University of North Carolina at Charlotte, Charlotte, NC
Email: {adoulat, iobaidat, minwoo.Lee}@uncc.edu

[1]Department of Software and Information Systems
[2,3]Department of Computer Science

***Abstract*— A vast amount of unstructured text that contains valuable information is available over the web. This text is changing and proliferating, making it hard for people to process, read, and remember. Data mining and information extraction algorithms are used to develop new automation techniques to process the unstructured text. Among this publicly available text, there are a considerable amount of online medical articles, which provides valuable information about diseases, symptoms, operations, treatments, drugs, etc. Automatic unstructured text classification offers practical information management that does not depend on the subjective criteria of classification. It also provides useful information by obtaining and correlating relevant data present in documents. It also classifies, identifies and presents all sources of knowledge and reduces the time for retrieving information by simplifying access to content. Therefore, medical information needs to be classified into their respected categories (such as Diabetes, Cancer, Depression, Pediatrics, etc.). In this paper, we propose to use a deep learning approach for unstructured medical text classification at the document level. In our classification model we used two types of features: (i) content-based features (stylistic and complexity), and (ii) health domain-specific features. Moreover, rather than dealing with binary classification, this work handles multi-classes medical articles classification. This classification is done based on linguistic features that are extracted from the text, it also incorporates medical domain-specific terms/keywords as part of the classification feature set. These domain-specific features are extracted by applying topic modeling technique to spot the most probable terms for each medical class. Our experiments shows a reasonable classification accuracy for such a large number of classes.**

***Index Terms*— Machine Learning, Deep Learning, Natural Language Processing, Linguistic Analysis, Medical Text Classification, Topic Modeling.**

I. INTRODUCTION

The ever-increasing deployment of the Internet technology resulted in an enormous increment in the number of electronic documents that are available online. The availability of this massive unstructured text makes the automated text classification task very important [1]. Text classification Automation is a primary task in the field of Natural Language Processing (NLP). It is mainly used to assign the text documents to proper classes based on their content [2]. The publicly available documents exist in many different domains that exhibit different challenges and solutions because of its nature. Generally, the classification of different documents is applied to address different purposes such as sentiment classification [3], [4], [5], web pages classification [6], author identification [7], spam email filtering [8], and unstructured text classification [9], [10].

Traditionally, to classify a document, bag-of-words approach [11] is used to extract features to be used in supervised classification algorithms such as Support Vector Machine (SVM) [12] or Naive Bayes [13]. However, this approach has some drawbacks including neglecting the words order, and when the size of training data is small, it suffers from data sparsity. To overcome these limitations, recent studies focused on more complex features, for instance Hughes et al. [14] proposed a medical text classification model using Convolutional Neural Netwoeks (CNNs). In their study, they used more complex techniques to represent the classification features such as word2vec and doc2vec. However, in this study they classified medical data on the sentence level unlike our work which classifies medical data on the document level.

Medical data is available online in a significant amount. Such data provides valuable information about diseases, symptoms, operations, treatments, drugs, etc. These medical documents need to be classified into their respected classes (such as Diabetes, Cancer, Depression, Pediatrics, etc.) to obtain useful knowledge out of them. The classification process is a significant step towards further implementation of useful medical applications. For instance, designing automated medical diagnosis tools or automated medical treatment tools [15]. However, the majority of the online medical information is not classified, which makes it hard to obtain useful knowledge out of it [16].

Few studies in the literature have addressed multi-class medical text. Rather most efforts focused on binary class problems [17], [18], [19]. On the other hand, most of previous studies are focused on handling medical text at the sentence level. For instance, questionnaire [20], social media posts and tweets [19], etc.

To help rectify this situation, we propose to use a deep learning approach that can help in classifying medical articles. The classification is based on features that can be extracted from these articles. These features include content-based features (stylistic and complexity), and health domain-specific features. The former have been used in the literature in many domains, such features by themselves are not sufficient enough to capture the target class of a given text (See Section II-A for more details). Domain-specific terms/keywords can play a major role in capturing the respected class of a given text. Topic modeling offers a methodology to capture the keywords and phrases correlated with each medical domain. Also, it is capable of associating a topic with a distribution over a set of words that represent the list of most probable words for each class. These keywords form a suitable set of features for the classification process.

To use the raw medical data in machine learning models, we need to prepare this data and extract the useful features for our model. This preprocessing process involves several steps including text sanitization, stop words/punctuation removal, sentence splitting, POS tagging, word tokenization, word lemmatization, and Named Entity Recognition (NER). We will discuss these steps in more details in Section III-C.

Given such excessive amount of raw medical data, the task of feature extraction is as critical as ever for the successful application of machine learning. Furthermore, the availability of many features makes it so hard to select the set of most significant features that can play a significant role in generating high accuracy for the classification problem. Feature selection is defined as the process of detecting relevant features and discarding irrelevant and redundant features with the goal of obtaining a subset of features that accurately describe a given problem with minimum degradation of performance [21]. Theoretically, having many input features might seem desirable, but some of these features might be uncorrelated with the target class which will affect the classification task negatively.

Specifically, this study will integrate deep learning techniques and linguistic analysis (Semantic and Statistical) techniques to automatically classify medical articles into their respected categories. There are four research questions that this study will answer:

- Whether using linguistic analysis results as input to deep learning models can perform well in medical articles classification.
- Whether incorporating domain specific terms (keywords) as a part of the classification features can improve the classification of the medical articles.
- How well the deep neural network models perform at classifying different medical article categories? Moreover, whether they can handle multi-class medical categories?
- Are deep neural network models useful in categorizing medical text in a large body of medical articles (document level classification) unlike previous studies that handle questionnaire, and social media posts?

The rest of this paper is structured as follows. Section II examines recent research in the field of text classification, section III presents our proposed methodology for medical text classification. On the other hand, it provides the details of our dataset, data prepreprocessing, classification feature extraction, and classification feature selection. Our proposed model evaluation is presented in section IV. Finally, we conclude and present our future plans in section V.

II. Related Works

This study brings together two threads from the literature: Linguistic analysis (Semantic and Statistical Models) and Deep learning Models.

Linguistics involves the analysis of language context, language meaning, and language form. The study of language semantics involves how to encode relations between properties, entities, and other aspects in the language to deliver, assign, process meaning, as well as to resolve and manage ambiguity [22].

Deep Learning (DL) is causing significant advances in finding solutions for problems that continued to be obstacles in the artificial intelligence area for many years. It enabled complex models that consist of multiple processing layers to learn data representation with many abstraction levels. These algorithms have improved the state-of-the-art in visual object recognition, speech recognition [23], NLP [24], and many other fields.

These two threads are deployed for the purpose of classification of unstructured medical text. In the following subsections, we will briefly review each of them.

A. Unstructured Text Classification

Bag of Words (BoW) is the most common method to describe text documents for classification and retrieval purposes. In BoW approaches, the description of a document is represented as a vector of term frequency of the document words [25]. Term Frequency/Inverse Document Frequency (TF/IDF) are the most prominent measurement of BoW [26].

The weighted term frequency vectors have a drawback; same topics documents might not be recognized as similar. This occurs if the used terms are not overlapping enough. Topic modeling was introduced to overcome this drawback. In this approach, the words are mapped to a latent representation space, where it describes the possible topics. The well known topic modeling is Latent Dirichlet Allocation (LDA) [27]. In LDA, sampling technique is used to refine topics iteratively. This process happens in a way that the resulting model produces a similar distribution of terms from a training document of same topic distribution.

Different classification approaches were presented in the text classification field. Teahan et al. [28], used a cross-entropy approach for the classification of text. This study is built on the basis that the entropy is one of the best methods for information content evaluation in a text data. A text classification system was developed by Kautz et al. [29] for multiple classes datatype. The ”imbalance” dataset was used for the analysis of their findings. They used 4-way ANOVA analysis for the feature selection in their study. This work suggests a new performance measure, named multi-class performance score (MPS) rather than using the well-known conventional measures such as the area under the curve (AUC) and receiver operating characteristic (ROC).

MPS had a minimum influence on the conditions of the testing and training data on all multiple class problems.

Other studies uses TF/IDF for the feature extraction. For instance, Boulis et al. [30] used unigram and bigram based features in their model. This approach does not affect the TF/IDF values. However, in their work, they increased the number of unique vocabulary features to increase the classifiers performance. Forman [31] used Bi-Normal Separation (BNS) instead of IDF for feature extraction. BNS uses distinguishing power to rank the terms. This study showed that scaling the importance of terms using BNS improved the accuracy of the classifier without any feature selection. Largeron et al. in [32] proposed an entropy-based approach. This approach is called Entropy-based Category Coverage Difference (ECCD). ECCD calculates the entropy of the terms in the classes to obtain the importance among different classes. Liu et al. [33] used a weighting scheme for terms based on the probability, this improved the performance of the classifier. Lu et al. [34] categorized biomedical data by a systematic modification of the TF/IDF scores. The SVM classifier performance was enhanced when the modified features were used. Another study used a modified version of TF/IDF which is called Delta TF/IDF [35]. Delta TF/IDF is a modification of the TF/IDF score, in which it is intended to integerate the sentiment score with the state-of-the-art TF/IDF. The difference between the negative and positive sentiment in the training data is used to calculate the delta value.

B. Medical Text Classification

Researcher attempted to classify unstructured text in the medical domain. McRoy et al. [20] proposed a classifier to answer community-based questions. The scheme of the classification includes a set of questions such as clinical, non-clinical, and patient-specific questions. Other efforts have been carried out to classify several aspects regarding medical field. Among these studies, Chomutare [36] used a classifier to predict patients with depression to provide help. Yang et al. [37] applied classification techniques to detect posts that discuss drug reactions. Tuarob et al. [17] used classification to detect if a twitter post is health related or not. Akbari et al. [38] used classification to identify wellness events, i.e. activities performed related to health, exercise, or diet. On the other hand, [18], [19] attempted to classify posts' authors in an online health community website. Whether an author of a post is a health professional or not. Abdaoui et al. [19] determined the authors of a post to be a lay man or health professional based on medical ontologies such as UMLS. Jagannatha and Yu [39] applied recurrent neural network sequence labeling in the detection of phrases in medical text. Kuo et al. [40] developed a natural language processing ensemble pipeline. This pipeline combines two systems MetaMap [41] and cTAKES [42] for extracting biomedical data element from clinical text. Then these biomedical elements were used for classification purposes.

The limitations of previous studies, can be generalized as follows:

- To the best of our knowledge, very few studies have been carried out for classification of multi-classes medical data. But rather, the majority of previous studies have focused on handling binary classes.
- Most of previous studies used the state-of-the-art approaches like BoW, TF/IDF for feature extraction. Although, these approaches performed well in some domains with small text excerpts, they perform poorly when dealing with large body of text because they have poor similarity values. Statistical analysis like topic modeling performs better in such scenarios.
- The majority of previous studies are focused on handling medical documents at the sentence level. For instance, questionnaire, social media posts, tweets, etc.

The key contributions of this work can be summarized as follows: (1) handling multi-classes medical data, while integrating medical domain specific terms (topics) as a part of the classification features, and (2) handling medical news articles at the document level unlike most of the previous studies that deal with small unit of text (sentences).

III. Methodology

To begin exploring the content of medical articles, we develop our own dataset of medical articles. In the following subsections, we present the steps of our proposed classification model. On the other hand, we discuss our dataset, data preprocessing, feature extraction/selection, and classification model selection. The steps of our proposed model are shown in Fig. 1. In this figure, we have: Fig. 1-a) The unstructured medical text, Fig. 1-b) The preprocessing steps we applied in our raw text, Fig. 1-c) The feature extraction and selection step, and Fig. 1-d) The model training on the selected set of features. In the following subsections, we discuss these steps in details.

A. Data Collection

To test our proposed model, we collected a dataset of about 100,000 medical articles. These articles were crawled from "Medical News Today: Health News website" [43]. They provide information about diverse diseases, symptoms, operations, and drugs. On the other hand, these articles are labeled by human experts into several categories including but not limited to: Diabetes, Cancer, Public health, Depression, Nutrition, Neurology, Breast cancer, Cardiology, and Infectious diseases. The categories for these articles are also crawled from "Medical News Today: Health News website".

B. Dataset Overview

Our dataset contains medical articles associated with ten different medical categories. These articles provide information about diseases, symptoms, operations, treatments, and drugs. On the other hand, the medical categories include "Diabetes", "Infectious Diseases / Bacteria / Viruses", "HIV / AIDS", "Cancer / Oncology", "Cardiovascular / Cardiology",

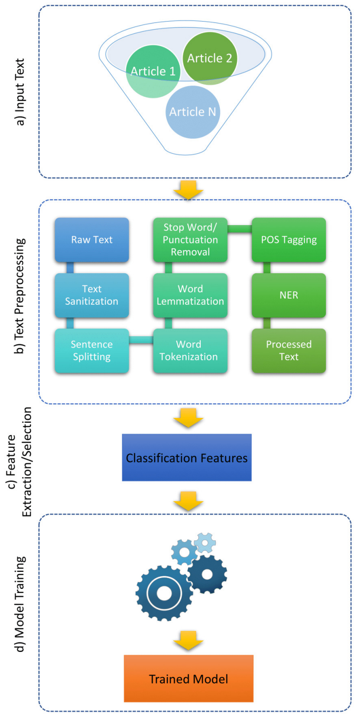


Fig. 1. Model Architecture

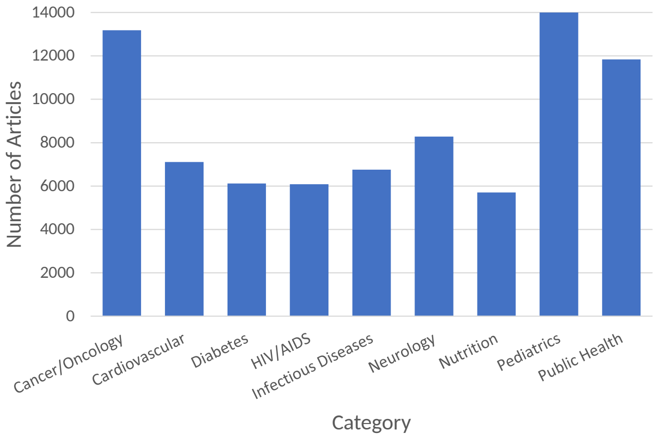


Fig. 2. Dataset Overview.

”Public Health”, ”Nutrition / Diet”, ”Neurology / Neuroscience”, and ”Pediatrics / Children’s Health”.

Fig. 2 shows the distribution of our dataset among different classes. Most of the classes are relatively balanced (e.g. Diabetes, Cardiovascular, HIV/AIDS, etc.). The class with the highest frequency is Pediatrics while Nutrition is the lowest. Since our dataset is relatively balanced, we do not need to perform undersampling or oversampling process.

C. Data Preprocessing

In this step, we preprocess the crawled articles to clean up our dataset. The preprocessing phase includes the following steps:

- Text Content Sanitization: in this step, we remove noise contents such as HTML tags. We then identify and filter out irrelevant content (such as script codes, advertisements, non-english words) that do not contain medical information.
- Sentence Tokenization: we use NLTK [44] sentence tokenizer to extract meaningful sentences from our dataset.
- Word Tokenization: we use NLTK word tokenizer to partition the text into a sequence of tokens, which roughly correspond to words. These words will be further used in the feature extraction process of our proposed model.
- Stop Words Removal: to filter out useless words. These stop words are commonly used words (such as ”the”, ”a”, ”an”, ”in” etc.). For this step, we used the predefined list of stop words by NLTK.
- Punctuation Removal: Again, we use NLTK word tokenizer to pick out sequences of alphanumeric characters as tokens and drop everything else (punctuation).
- Word Lemmatization: in this step, we use NLTK WordNet Lemmatizer. This aims to remove inflectional endings only and to return the dictionary form of a given word.
- Part Of Speech (POS) Tagging: For this task, we use NLTK POS tagger to extract part of speech tags (such as noun, verb, determinant, etc.).
- Named Entity Recognition (NER): Finally, in this step, we use the NLTK wrapper over the NER from Stanford CoreNLP to extract all textual mentions of the named entities (such as person, location, date, etc.).

The preprocessing steps that we have applied to our dataset are shown in Fig. 1-b.

D. Classification Features Extraction

To study different medical articles’ categories, we extract two types of features: content-based features and domain-specific features. In the following subsections, we describe these features in more details.

1) Content-Based Features: To spot the main differences between various medical articles’ categories, we extract content-based features from our dataset. These features can be classified into two main classes: stylistic and complexity features.

- Stylistic Features: these features are based on NLP to understand the text style, syntax, and grammatical components of each medical article content. For this purpose, we use NLTK Part Of Speech (POS) tagger and keep track of each tag frequency within an article. For instance, we count the number of nouns, verbs, proper nouns, determinants, comparatives and superlatives, etc. Along with this, we use the 2010 Linguistic Inquiry and Word Count (LIWC) dictionaries [45] to keep track of the frequencies of negation, belief, surprise, conditional, modal, existential, and interjection words. Additional stylistic features include a count of capitalized words, wh-words.

- Complexity Features: To capture the complexity of each medical class text, we exploit two levels of complexity: the sentence and the word levels. For the former one, we computed the average sentence length, the lexical diversity (number of unique words to the total number of words in a given sentence). For the latter one, we compute the average word length and count the number of words with length more than a specific threshold. Additional features regarding word level complexity are based on the readability of a medical article. For this purpose, we use a Python package called textstat [46] to calculate statistics from text to determine readability, complexity and grade level of a given article. This package includes three different grade level readability measures: Gunning Fog, SMOG Grade, and Flesh-Kincaid grade level measure. These measures use the number of syllables in a word to compute a grade level reading score for this word. We assume that a higher score means a medical article needs a higher education level to read.

2) Domain-Specific Features: To improve the performance of our classification model, concerning prediction accuracy, precision, and recall. We propose to extract domain-specific terms (keywords) for each medical class. For this purpose, we use LDA topic modeling on a subset of articles from our dataset. The central intuition behind this is to mine the keywords and phrases correlated with each medical class. Since, LDA is based on a (generative probabilistic) model that associates a topic with a distribution over a set of words (and, as a corollary, each topic has its list of most probable words). Therefore, we create a set of keywords/phrases associated with each medical class. Then we exclude words with probability less than a specific threshold, to eliminate words that are less-domain specific. As a result of this, we have a set of keywords for each class in our dataset, each set represents a single feature. However, the subset of articles that we use for the domain-specific keywords extraction is excluded from further model training and testing.

E. Classification Features Selection

After the feature extraction phase, we used the Recursive Feature Elimination (RFE) [47] with cross-validation to select the set of most significant features (most correlated features with the target class). RFE automatically tune the number of features selected with cross-validation. Fig. 3 presents the cross validation score (in terms of the number of correct classifications) for different combinations of features. In this figure, the optimal number of features that produces the best classification accuracy is 42 out of our 58 extracted features. We applied RFE for two main reasons:

- To Reduce the number of classification features, to avoid overfitting and to ensure the generalization of our model.
- To obtain a better understanding of the set of selected features and their direct relationship to the target class.

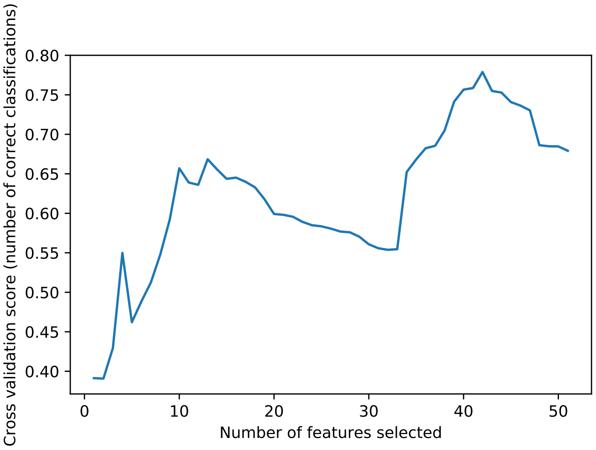


Fig. 3. Feature Selection using RFE

IV. Model Evaluation: Accuracy, Precision, and Recall

In this section, we evaluate our proposed model in terms of prediction accuracy, precision, and recall. We trained a deep neural network model on the set of extracted features. During the evaluation process, we tested various deep network configurations. We experimented with multiple configurations of network, in terms of the number of hidden layers including 2, 3, 4, 5 and 6 layers. We also experimented with a different number of neurons per hidden layer including 20, 50, 70 and 100 neurons. Moreover, we tired several activation functions including *no-op activation*, *logistic sigmoid*, *hyperbolic tan*, and *rectified linear unit*. Also, we tried several weight optimization solvers including *sgd*, which refers to the stochastic gradient descent, *adam*, which is a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba, and *lbfgs*, which is an optimizer in the family of quasi-Newton methods. We also tried several values for the regularization term α including 0.001, 0.01, and 0.1. These experiments were conducted to identify the optimal neural network configurations, the resulted optimal architecture of the deep neural network is illustrated in Fig. 4.

TABLE I

The feature sets and their obtained accuracy.

Features Set	**Accuracy**
Stylistic Features	28%
Complexity Features	13%
Domain-specific Features	49%
Combined Features	74%
Optimal Set Features	82%

We trained our model using different sets of features: (i) Stylistic Features, (ii) Complexity Features, (iii) Domain-specific features, (iv) A combined set of all previous features, and (v) Optimal set of features. The classification results for each set of features using MLP approach are depicted in Tab. I. It is obvious from the table that the domain-specific features produce the highest accuracy among the feature sets. After combining all of the extracted features, and selecting

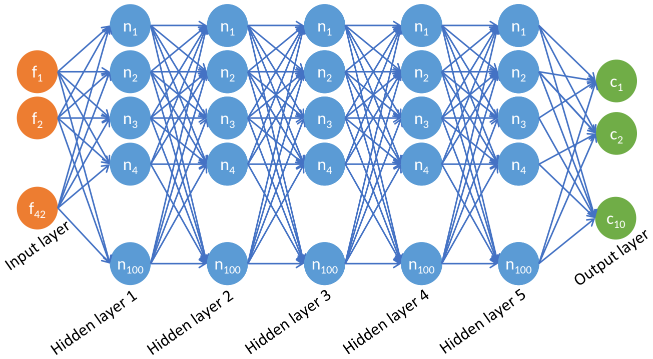


Fig. 4. Neural Network Architecture: The optimal neural network configurations for the proposed model.

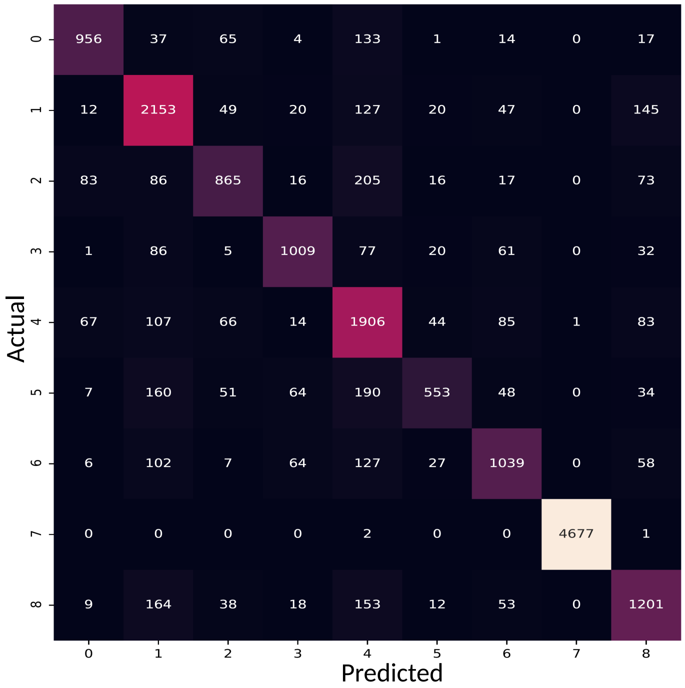
the optimal set of features (using RFE), we performed 10-fold-cross-validation on this set, and the resulting accuracy is 82%. The precision, recall, and F1-score for our model were 81, 82, 80 respectively. We compared our obtained accuracy with a baseline model. We used the-state-of-the-art TF/IDF for this purpose, and we obtained 62% accuracy result. Therefore, our model outperforms the baseline model by 10%.

Fig. 5 shows the confusion matrix of the classification results. This figure shows the discrepancies between the predicted and actual labels, we can notice that The vast majority of the predictions end up on the diagonal, where the predicted labels equal to the actual labels. Another observation from the figure is that the proposed model performs best in predicting the pediatrics class, while public health has the lowest prediction results. The reason for the low score of public health is that it has a lot of common terms that can be in any medical class, and this proves that the domain-specific features play a significant role in the classification process.

V. CONCLUSIONS

Unstructured text classification has gained a world wide attention in the past couple of decades. It is an important step in NLP towards further analysis on unstructured text. In this work, we investigate the classification of online medical articles using linguistic (Semantic and Statistical) analysis. On one hand, this work is an attempt to distinguish the different writing styles and complexities between different medical domain specialists. On the other hand, this work shows that incorporating domain-specific terms and keywords can effectively improve the classification accuracy of the machine learning models when it comes to specific domains. We also trained a deep neural network on our proposed set of features to be used for medical articles classification. The obtained accuracy using our proposed set of features is 82% which is higher than the accuracy using the baseline TF/IDF model which is 62%.

This work is the first step towards further designing health-related applications. After identifying the different classes for the medical articles, we plan to extract symptoms related to each class (disease), possible treatments for such symptoms, etc. We will address these issues in future work.



	0	1	2	3	4	5	6	7	8
0	956	37	65	4	133	1	14	0	17
1	12	2153	49	20	127	20	47	0	145
2	83	86	865	16	205	16	17	0	73
3	1	86	5	1009	77	20	61	0	32
4	67	107	66	14	1906	44	85	1	83
5	7	160	51	64	190	553	48	0	34
6	6	102	7	64	127	27	1039	0	58
7	0	0	0	0	2	0	0	4677	1
8	9	164	38	18	153	12	53	0	1201

Fig. 5. Model Accuracy

REFERENCES

[1] B. Zaqaibeh, I. Obaidat, and W. Hussien, "Big data analysis techniques using multi-gpus mapreduce implementations," *International Journal of Advanced Studies in Computers, Science and Engineering*, vol. 5, no. 11, p. 1, 2016.

[2] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[4] I. Obaidat, R. Mohawesh, M. Al-Ayyoub, A.-S. Mohammad, and Y. Jararweh, "Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, pp. 1–6, IEEE, 2015.

[5] M. Al Smadi, I. Obaidat, M. Al-Ayyoub, R. Mohawesh, and Y. Jararweh, "Using enhanced lexicon-based approaches for the determination of aspect categories and their polarities in arabic reviews," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 11, no. 3, pp. 15–31, 2016.

[6] S. A. Özel, "A web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3407–3415, 2011.

[7] C. Zhang, X. Wu, Z. Niu, and W. Ding, "Authorship identification from unstructured texts," *Knowledge-Based Systems*, vol. 66, pp. 99–111, 2014.

[8] I. Idris, A. Selamat, N. T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 33–44, 2015.

[9] A. J. J. Yepes, L. Plaza, J. Carrillo-de Albornoz, J. G. Mork, and A. R. Aronson, "Feature engineering for medline citation categorization with mesh," *BMC bioinformatics*, vol. 16, no. 1, p. 113, 2015.

[10] V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with laplacian svms: an application to cancer case management," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 869–875, 2013.

[11] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[13] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[14] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," *Stud Health Technol Inform*, vol. 235, pp. 246–50, 2017.

[15] H. C. Koh, G. Tan, *et al.*, "Data mining applications in healthcare," *Journal of healthcare information management*, vol. 19, no. 2, p. 65, 2011.

[16] S. Ananiadou, D. B. Kell, and J.-i. Tsujii, "Text mining and its potential applications in systems biology," *Trends in biotechnology*, vol. 24, no. 12, pp. 571–579, 2006.

[17] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram, "Discovering health-related knowledge in social media using ensembles of heterogeneous features," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 1685–1690, ACM, 2013.

[18] J. Huh, M. Yetisgen-Yildiz, A. Hartzler, D. W. McDonald, A. Park, and W. Pratt, "Text classification to weave medical advice with patient experiences.," in *AMIA*, 2012.

[19] A. Abdaoui, J. Azé, S. Bringay, N. Grabar, and P. Poncelet, "Predicting medical roles in online health fora," in *International Conference on Statistical Language and Speech Processing*, pp. 247–258, Springer, 2014.

[20] S. McRoy, S. Jones, and A. Kurmally, "Toward automated classification of consumers cancer-related questions with a new taxonomy of expected answer types," *Health informatics journal*, vol. 22, no. 3, pp. 523–535, 2016.

[21] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.

[22] A. Martinet and E. Palmer, "Elements of general linguistics," 1966.

[23] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[24] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

[25] R. R. Larson, "Introduction to information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 4, pp. 852–853, 2010.

[26] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, pp. 305–338, Nov 2016.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[28] W. J. Teahan and D. J. Harper, "Using compression-based language models for text categorization," in *Language modeling for information retrieval*, pp. 141–165, Springer, 2003.

[29] T. Kautz, B. M. Eskofier, and C. F. Pasluosta, "Generic performance measure for multiclass-classifiers," *Pattern Recognition*, vol. 68, pp. 111–125, 2017.

[30] C. Boulis and M. Ostendorf, "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams," in *Proc. of the International Workshop in Feature Selection in Data Mining*, pp. 9–16, Citeseer, 2005.

[31] G. Forman, "Bns feature scaling: an improved representation over tf-idf for svm text classification," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 263–270, ACM, 2008.

[32] C. Largeron, C. Moulin, and M. Géry, "Entropy based feature selection for text categorization," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, pp. 924–928, ACM, 2011.

[33] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," *Expert systems with Applications*, vol. 36, no. 1, pp. 690–701, 2009.

[34] X. Lu, B. Zheng, A. Velivelli, and C. Zhai, "Enhancing text categorization with semantic-enriched representation and training data augmentation," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 526–535, 2006.

[35] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis.," *Icwsm*, vol. 9, p. 106, 2009.

[36] T. Chomutare, "Text classification to automatically identify online patients vulnerable to depression," in *International Symposium on Pervasive Computing Paradigms for Mental Health*, pp. 125–130, Springer, 2014.

[37] M. Yang, M. Kiang, and W. Shang, "Filtering big data from social media–building an early warning system for adverse drug reactions," *Journal of biomedical informatics*, vol. 54, pp. 230–240, 2015.

[38] M. Akbari, X. Hu, L. Nie, and T.-S. Chua, "From tweets to wellness: Wellness event detection from twitter streams.," in *AAAI*, pp. 87–93, 2016.

[39] A. N. Jagannatha and H. Yu, "Structured prediction models for rnn based sequence labeling in clinical text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016, p. 856, NIH Public Access, 2016.

[40] T.-T. Kuo, P. Rao, C. Maehara, S. Doan, J. D. Chaparro, M. E. Day, C. Farcas, L. Ohno-Machado, and C.-N. Hsu, "Ensembles of nlp tools for data element extraction from clinical notes," in *AMIA Annual Symposium Proceedings*, vol. 2016, p. 1880, American Medical Informatics Association, 2016.

[41] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[42] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[43] "Medical news today: Health news." https://www.medicalnewstoday.com/. Accessed: 6-5-2018.

[44] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 31, Association for Computational Linguistics, 2004.

[45] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[46] M. Huning, "Textstat 2.7 users guide," *TextSTAT, created by Gena Bennett*, 2007.

[47] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.