Early Detection of Severe Flu Outbreaks using Contextual Word Embeddings

Redouane Karsi[1], Mounia Zaim[2], Jamila El Alami[3]
LASTIMI Laboratory, Higher School of Technology of Sale
Mohammed V University
Rabat, Morocco

***Abstract*—The purpose of automated health surveillance systems is to predict the emergence of a disease. In most cases, these systems use a text categorization model to classify any clinical text into a category corresponding to an illness. The problem arises when the target classes refer to diseases sharing multiple information such as symptoms. Thus, the classifier will have difficulty discriminating the disease under surveillance from other conditions of the same family, causing an increase in misclassification rate. Clinical texts contain keywords carrying relevant information to distinguish diseases with similar symptoms. However, these specific words are rare and sparse. Therefore, they have a minor impact on machine learning models' performance. Assuming that emphasizing specific terms contributes to improving classification performance, we propose an algorithm that enriches training samples with terms semantically similar to specific terms using the deep contextualized word embeddings ELMo. Next, we devise a weighting scheme combining chi-square and semantic scores to reflect the relatedness between features and the disease under surveillance. We evaluate our model using the SVM algorithm trained on i2b2 dataset supplemented by documents collected from Ibn Sina hospital in Rabat. Experimental results show a clear improvement in classification performance than baseline methods with an F-measure reaching 86.54%.**

Keywords*—ELMo; SVM; contextual word embeddings; semantic term weighting; health surveillance; text classification

I. Introduction

Public health surveillance is a significant focus of National health policies. It is ensured by collecting epidemiological data from various healthcare facilities to detect disease outbreaks and subsequently plan appropriate response strategies early.

In Morocco's epidemiological surveillance system, the law requires healthcare producers to report all confirmed cases of notifiable diseases. For this, physicians must fill a particular form with the patient's clinical and demographic data. However, many physicians do not respect this notification formality, especially in the private sector. Thus, the total amount of collected forms is not entirely significant for correctly estimating epidemiological trends. Besides, the notification procedures for disease and collected data processing are not automated for acquiring relevant epidemiological indicators in real-time.

Today, several hospitals in the country have implemented the electronic health record (EHR), which appears to be an excellent opportunity for a better epidemiological surveillance system, because patient information captured and stored in EHR are so relevant for healthcare decision making [1], [2].

In EHRs, data is captured in a structured format, such as administrative data. Simultaneously, there is unstructured data written in a free text by practitioners. This textual data reflects the patient’s health status and helps determine exciting health indicators [3].

Text classification algorithms select meaningful information from EHRs to organize textual documents into a set of pre-defined categories [4]. In outbreak detection, a class corresponds to one disease.

Feature selection is a crucial element in the preprocessing phase. Its role is to optimally reduce feature space's dimensionality by selecting a subset of relevant terms according to some criteria [5].

The biggest challenge of feature selection methods is to correctly select features with high discriminative power. For this purpose, some methods rely on Frequency-based feature selection[6], [7], others like Information Gain (IG) and chi-square test rank terms according to their correlation with the class variable [8], [9]. More recently, a new research trend favors semantic similarity based on knowledge resources and the fast-growing field of deep neural networks [10].

The flu surveillance system's goal is to predict the spread of a severe form of influenza accurately. The acquired flu-related free-text clinical records are classified into two categories (severe flu or mild flu). These two forms of flu share many signs and symptoms. In this situation, the risk of misclassification increases, especially for documents related to severe influenza cases, since the frequency of specific features that characterize severe cases is low compared to common features frequency.

In this respect, many research efforts attempt to improve feature selection algorithms by highlighting the discriminative power of infrequent specific terms. Thus, ontology-based feature selection methods like UMLS and SNOMED CT have been intensively experimented with real improvements in the medical domain. [11], [12].

Despite the progress achieved in utilizing ontology-based feature selection methods, it is not sure that they are useful in differentiating between two similar classes as long as they share many terms. We can mention two reasons:

1) The clinical note: "The patient with fever, cough, and runny nose, diagnosed as positive for H1N1", reveals a severe case of flu, yet the three common terms in the document (fever, cough, and runny nose) are preponderant compared to the only specific term (H1N1) which despite its importance, it is very infrequent in the corpora, which might reduce classifiers efficiency [13].

2) We usually calculate feature weights according to their frequencies or their statistical correlation with the target class. However, the rare term "H1N1" can be underestimated despite being semantically more heavily weighted than all other features.

To overcome the shortcomings of statistical and ontology-based feature selection methods, static word embeddings models have been put forward because of their ability to capture word semantic proprieties [14]. However, they are inefficient in handling the widely varying medical spelling since they provide a unique word representation. Hence, we hypothesize that a contextual word embeddings representation [15] with a weighting scheme combining statistical and semantic scores can better emphasize rare medical words improving classification performance in outbreaks detection.

In this paper, we propose in a first step an algorithm that aims to enrich training samples related to severe cases of flu with features that are semantically similar to specific features. The idea behind this algorithm is to mitigate the deficiency caused by the scarcity of specific features by adding new features to training samples in order to counterbalance the preponderance of common features. This algorithm is based on a deep contextualized word representation method named: Embeddings from language models (ELMo), renowned for its power in detecting the finest syntactic and semantic characteristics of words. In a second step, the weight of specific terms is determined by combining two measures: The chi-square weight calculated from the information class provided by labeled data and the semantic weight that corresponds to a score assigned considering the term’s association with a severe respiratory illness.

We evaluate the proposed feature selection model using SVM Classifier and the clinical dataset i2b2 [16] enriched with clinical reports gathered from the EMR of the Ibn Sina hospital in Rabat. Experimental results show significant improvement compared to ontology-based feature methods and static word embeddings models with a notable decrease in misclassification rate of test clinical notes related to severe flu by reaching an F-measure of 86.54%.

The principal contributions of this work are: firstly, a novel approach to extend rare and high discriminative words using a contextual word embeddings model. Secondly, a new weighting scheme combining a statistical and a semantic score.

In the remainder of this paper, an overview of related work is provided, followed by a description of our feature engineering approach. Then experimental results are presented and discussed. In the last section, we conclude our work.

II. Related Work

Traditional health surveillance systems rely on epidemiological data collected periodically from various public health system bodies to detect the appearance of a disease [17]. With the emergence of social media and the progressive use of EMRs in healthcare facilities, much health-related data is becoming available to feed automated health surveillance systems with relevant data. In the literature, different approaches have been proposed to take advantage of the textual information available in health-related documents to develop efficient disease prediction systems.

Concerning statistical approaches, SVM, n-gram features, and negation algorithm (NegEx) are experimented in [18] to predict diagnoses from intensive care unit notes. They found that bigrams perform better than other n-gram representations. The negation algorithm does not improve the performance of unigrams. In the study described in [19], the death certificates are classified by type of cancer-causing death. The n-grams and features extracted from the SNOMED CT ontology are employed together to train an SVM classifier, except for certain rare and ambiguous cancers, the proposed model remains effective. In the two previous studies, the authors reported that their classification models are resource-intensive and time-consuming. To overcome this defect, feature selection methods are adopted because they help select relevant features while reducing feature space dimension.

A feature selection approach based on chi-square and t statistical tests is proposed in [20]. It consists of selecting a ranked subset of features from different intensive care unit reports. A configurable threshold determines features list size. A binary classification model is then trained on n-grams, UMLS concepts, and assertion values associated with pneumonia expressions as features to identify pneumonia cases. Experiments show that the number of selected features has no significant impact due to noisy features. In terms of performance t-test, the union of t and chi-square tests and the combination of all feature types provide the best results.

Motivated by the breakthrough of ontologies in the medical domain, many researchers are working to exploit the knowledge presented in ontologies to make predictions on events related to the medical domain. For example, the extended syndromic surveillance ontology is developed in [21]. Its role is to facilitate early disease prediction. It is designed to identify clinical text concepts and then associate the extracted concepts with a particular syndrome. This ontology is created around concepts and their relations, which is tedious and requires excellent domain expertise. Moreover, automated ontologies are conceived in [22] when the proposed model inferred new relations between medical concepts discovered in clinical texts. For this, the model finds its strength in using linked biomedical ontologies to extract relations from enriched concepts.

Despite their power, ontologies are very expensive to setup, because they require domain expertise and are based on standard terminology that changes very little. Therefore they do not take advantage of the explosion of knowledge-rich textual content encapsulated in linguistic forms. An exciting alternative is to use deep neural networks to learn word embeddings to generate a semantic representation of words in a vector space. In this way, the semantic similarity between words will be determined only by a simple vector computation. Feature selection approaches using neural networks are considered to be useful in selecting the more relevant features. To illustrate this point, a hybrid feature selection method is described in [10] to infer the population's influenza rate. For this, the terms strongly correlated with the target concept are selected from the labeled data. Then, the word2vec language model generates word embeddings for the selected features and retains those with high similarity with the target concept. The problem of rare and out of vocabulary words is addressed in [23], a biomedical word embedding is created by exploiting the subword information. Word embeddings are good at enriching the terminology of existing concepts. They are used in [24] to extend the terminology of dietary supplements. The experimental results prove that the expanded terms are more relevant as search keywords in clinical notes than in external knowledge sources.

Term weighting is an essential step in the classification process. It aims to emphasize useful terms that contribute to better classification accuracy. Traditional methods such as TF-IDF have long proven their effectiveness. Thus, the work presented in [25] elucidates that the use of word2vec word embeddings with TF-IDF is effective for disease classification. In more recent studies, a semantic weight is suggested to express domain relatedness between concepts in the medical domain. We can cite as an example the research work discussed in [26], where word embeddings of all medical concepts are extracted from a corpus of biomedical texts. Then, an association score between each pair of concepts is calculated so that the weight of a concept in a document corresponds to the addition of its TF-IDF frequency with the sum of the association scores of its co-occurring concepts highly associated with it. The proposed weighting scheme outperforms the baseline TF-IDF.

In the literature, several feature-engineering techniques have been proposed. However, to the best of our knowledge, no existing searches explicitly address an approach that emphasizes rare discriminative words in the medical domain. Our work's novelty lies in using a contextual word embeddings model to extend rare features and a new weighting scheme combining statistical and semantic measures.

III. Our Feature Engineering Approach

To alleviate the problem of misclassification when the target classes share several common features, we present a feature enrichment method based on deep neural networks in conjunction with a term weighting scheme in order to strengthen the discriminative power of specific features contained in free-text clinical data. Our approach includes the following steps: Text preprocessing, specific features extraction, word embeddings generation and features weighting. The proposed model is depicted in Fig. 1.

A. Text Preprocessing

Clinical text is full of unnecessary and misspelled words that provide no added value, so before considering feature selection, the text was cleaned up by performing the following actions:

- Text tokenization: Consists of splitting the text into words.
- Text normalization: Consists of representing a word in its canonical form, for example, the words "went" and "going" will be normalized into the word "go".
- Stopwords removal: Words such as prepositions and articles are very common in documents but do not bring useful information for classification.
- Correcting misspelled words: As the dataset contains documents written in French, they are submitted to a spell checker before being translated into English.

B. Medical Concept Extraction

After eliminating stop words, documents are represented by terms that didn't have the same degree of relevance to discriminate between classes. Medical terms are more informative than non-medical terms, but a medical term can be expressed differently depending on the terminology practised by physicians. For example, the rise in body temperature can be designated by one of the terms (Fever, high temperature, hyperthermia). Therefore, a useful term is penalised by the problem of sparsity which reduces its discriminative power. To remedy this problem, the extraction of medical concepts plays a very important role in the normalisation of medical terms into a single synonym denser in documents.

In the context of this work, the MetaMap [27] program developed by the U.S. National Library of Medicine is used to extract medical concepts, its role is to parse the content of a biomedical text in order to recognize medical terms that refer to UMLS concepts.

UMLS organizes the concepts by semantic type, this structure of concepts provided by UMLS is helpful to exclude useless semantic types for prediction, thereby, in consultation with two physicians, we opted for the following semantic types deemed meaningful to predict diseases: Functional Concept, Finding, Virus, Sign or Symptom, Disease or Syndrome, Organic Chemical, Pharmacologic Substance, Medical Device. In Table I, an example of concepts extracted from a clinical text with their semantic types.

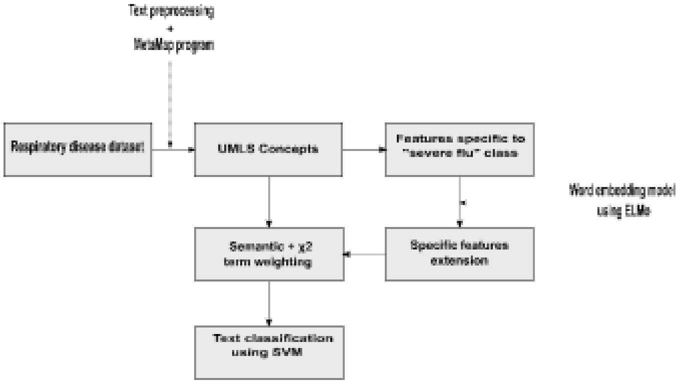
Text preprocessing
+
MetaMap program
Respiratory disease dataset
UMLS Concepts
Features specific to "severe flu" class
Word embedding model using ELMo
Semantic + χ2 term weighting
Specific features extension
Text classification using SVM

Fig. 1. Architecture of our Feature Engineering Approach.

- **Clinical text:** *"This is a 68-year-old patient with a history of **Type 2 Diabetes** under **ADO** and **chronic smoking** estimated at 30 PA, Admitted for the management of her **respiratory distress** attributed to **pneumonia**".*

TABLE I. EXAMPLES OF UMLS CONCEPTS EXTRACTED FROM A CLINICAL TEXT

Extracted concept	**Semantic type**
Type 2 Diabetes, Pneuminia	Disease or Syndrome
ADO	Pharmacologic Substance
Chronic smoking	Finding
respiratory distress	Sign or Symptom

C. Word Embeddings Generation

Concept extraction only partially solves the problem of sparsity, medical concepts are also written through abbreviations, acronyms and coding conventions of medical terms specific to each health system that cannot be mapped to UMLS. Furthermore, traditional classifiers cannot retrieve semantic similarities of rare words as they occur in few documents, and therefore have a minor contribution to classification accuracy. In this work, textual documents are to be classified into similar categories, in this case, specific terms are clearly in a minority compared to common terms, yet they are very decisive to predict severe flu. Thus, in order to emphasize these specific terms, documents will be extended with terms similar to them.

Word embedding is a recent technique powered by continuous advancements in deep learning, it is used to learn word vector representation to capture semantic properties helpful to quantitatively estimate the similarity between words.

To optimize the semantic representation of words, word embeddings are often pre-trained on large datasets so that words that occur in the same context have similar meanings.

Word2vec [28] and Glove [29] are among the best known methods, they are called static methods because they only produce a single representation of a word regardless of the context in which the word appears. For example, static methods generate the same representation of the word fever even if its meaning differs depending on the context.

- American election **fever** is approaching.
- **Fever** is a sign of Covid-19.

In this work, Elmo [30] is used to extend the terminology of medial concepts by capturing only the medical meaning of polysemous words. ELMo is a product of Allen NLP, its operating principle is based on two tasks: First, a deep bidirectional LSTM-based language model is pre-trained on a large textual dataset, then, in a second step, the hidden internal states of the model are used to generate the vector representation of words taking into account the context in which the word appears. ELMo can capture the finest syntactic and semantic aspects, and thus outperforms classic models like word2vec et GloVe in many NLP tasks.

In practice, a pre-trained ELMo model trained on PubMed [31] is used to generate 1024-dimensional embeddings of specific terms collected from the training data annotated as "severe flu". For this, it is necessary to identify specific terms that will be submitted to the model, thus, we have defined a specific term as a term whose frequency in the training documents labeled as "severe flu" is clearly higher than its frequency in the training documents labeled as "mild flu", this definition can be formulated as follows:

$$specificTerms = \left\{ t \in Tsf, \frac{D(t,Cmf)}{D(t,Csf)} \leq \alpha \right\} \qquad (1)$$

Where:

Csf is the class label of training documents related to "severe flu".

Cmf is the class label of training documents related to "mild flu".

Tsf is the training set of class "severe flu".

D(t, Csf) is the number of documents of class Csf containing the term t.

D(t, Cmf) is the number of documents of class Cmf containing the term t.

α is the threshold which determines the list of specific terms to be selected.

Since ELMo may generate multiple embeddings per word, we average the vectors of all the occurrences of each term to obtain the corresponding word vector.

The i2b2 dataset word vectors are associated with specific term embeddings to form a base of eligible words to extend terminology for specific words. A cosine similarity-based measure is calculated between each specific word and all eligible words, so that except words that reach a similarity above a certain threshold will be retained. In Table II, we list some specific words with their closest similar concepts with a threshold equal to 0.75.

TABLE II. SOME UMLS CONCEPTS WITH THEIR CORRESPONDING MOST SIMILAR WORDS

UMLS concept	**Most similar words**
Pneumonia	Dyspnea, Desaturation, cyanosis, tachypnea
Swine	H1N1, virus, flu, SRAS, coronavirus, pandemic
distress	ARDS, respiratory, breath, dyspnea
Intubation	Ventilation, nebulization, respirator, ICU
diabetes	Sugar, insulin, hyperglycemia

Although the word *"Swine"* refers to the animal domain, its embedding generated by the model is closely related to the medical context, the same for the word *"Distress"* which is encountered in several contexts but attributed to the medical domain.

D. Term Weighting Scheme

The proposed weighting scheme attempts to assign an appropriate weight to each extracted term and the list of extended features based on their power to discriminate between severe and mild flu. The frequency-based weighting scheme is not convenient due to the fact that we have at our disposal a labeled training set that tells us about the degree of correlation between terms and the target classes, in addition, even if two terms have a strong correlation with the class "Severe flu", it may not have the same semantic importance for the target category, for example, the term "Pneumonia" is more indicative of a severe case of flu than the term "Fever". Thus, it makes more sense for the proposed weighting scheme to take into account both the degree of correlation with the target class and the semantic importance of terms.

The chi-square test is performed to test the hypothesis of independence between two categorical variables. In text classification, this test is used to rank words from a corpus of textual documents in order to select those that strongly depend on the target class. This dependence between variables is measured by the chi-square test by applying the formula below.

$$\chi^2(d,f,c) = \sum_{i=o}^{1} \sum_{j=0}^{1} \frac{(O_{ij}-E_{ij})}{E_{ij}} \qquad (2)$$

Where

O and E are respectively, the observed and expected numbers.

The index i indicates whether the term f is present or not in document d.

The index j indicates whether the document d belongs to class c.

A greater value of the chi-square test indicates that there is a strong correlation between the term and the corresponding class, for that, we retain the chi-square value to calculate the weight of words.

In addition to the weight generated by the statistical correlation between words and the target class, a weight reflecting the semantic importance is used to determine the final weight of a word so as to assign a high semantic weight to terms that are more indicative of severe flu, to do this, the semantic weight of a term corresponds to the cosine similarity between the term and the class "Severe flu", and in order to simplify the calculation of similarity, the class "Severe flu" is represented by the word *"Pneumonia"* since it is often associated with severe complications of flu. In short, the semantic weight is formulated as follows:

$$semantic_w(t) = cosSimilar(emb(t), emb(pneumonia)) \qquad (3)$$

Finally, the final weight of the term t is calculated by associating the chi-square weight with the semantic weight according to the formula below:

$$finalWeight(t) = \beta.semantic_w(t) + (1-\beta).\chi_w^2(t) \qquad (4)$$

Where

$semantic_w(t)$ is the semantic weight of the term t.

$\chi_w^2(t)$ is the chi-square weight of the term t.

β is a parameter which determines the share of semantic weight in the final weight.

emb(t) is the vector word representation of the term t.

IV. Results and Discussion

A. Datasets

To conduct our experiment, two sources were used to collect clinical documents pretaining to flu.

1) For severe flu cases: By searching with the keyword "Pneumonia", several clinical notes were extracted from the i2b2 dataset, then with the support of two physicians, 500 documents are labelled as "severe flu".

2) For mild flu cases: We collected and translated into english reports of all medical consultations carried out by the pneumology department of Ibn Sina hospital in rabat during the period between January 2017 and February 2020 provided that these consultations did not result in hospitalization. Among the collected documents, 500 reports were annotated as "mild flu".

In the remainder of this section, "Severe flu" documents are considered as positive samples, while "Mild flu" documents are regarded as negative.

B. Evaluation

A health surveillance system must be efficient enough to accurately detect the onset and progression over time of a disease, so our proposed model is designed to meet the following two requirements:

1) Reduce the proportion of mild flu-related documents classified as severe flu, this has the effect of avoiding false outbreak alerts. To assess this performance, the precision which represents the percentage of correct positive predictions over positive predictions is measured as follows:

$$Precision = \frac{T_P}{T_P+F_P} \qquad (5)$$

2) Reduce the proportion of severe flu-related documents classified as mild flu in order to prevent an outbreak going undetected. This performance is evaluated through the recall measure which is defined as the percentage of correct positive predictions over the total number of actual positive documents, and it is calculated as below:

$$Recall = \frac{T_P}{T_P+F_N} \qquad (6)$$

The proposed classification model is considered sufficiently perfect once high values of precision and recall are reached. In our model, finding a good compromise between precision and recall amounts to determining the values of these two measures which maximize their harmonic mean, also called F-measure. The F-measure is calculated as follows:

$$F-measure = 2.\frac{Precision.Recall}{Precision+Recall} \qquad (7)$$

Where

T_P is the number of true positive predictions.

F_P is the number of false positive predictions.

F_N is the number of false negative predictions.

C. Experimental Results and Discussion

Our feature engineering approach is confronted with three baseline methods:

1) Bag of words + TF-IDF: Words resulting from text preprocessing tasks are extracted, then weighted using TF-IDF.

2) UMLS concepts + TF-IDF: Clinical text is mapped to the UMLS concepts using the MetMap program, then the TF-IDF weight is calculated for each extracted concept.

3) The word embeddings model word2vec pretrained on PubMed.

The SVM method is applied to all types of features used in our experience with 90% of the dataset goes to the training set, and the remaining 10% to the testing set. SVM is an ideal choice, as it performs well on different domains [32].

The performance measures of our model depend on two parameters α and β explained in detail in the previous section.

1) α is the threshold below which a term is considered specific to the positive class (Severe flu).

2) β is the share of the semantic weight in the final weight of a term.

When the value of α is very small. The model extracts terms specific to the positive class which rarely occur in the negative class. On the other hand a large value of β means that the semantic weight of a term is greater than its statistical weight.

According to experimental results presented in Fig. 2 the recall value is at its lowest when the model allows selecting as specific term those whose frequency in the positive class is close to their frequency in the negative class, i.e. α close to 1, which can be explained by the fact that the model tends to extend the terminology of common terms that are already dominant over specific terms, causing an increase in FN.

When the value of α decreases, we notice a clear improvement in recall, because the model rejects more common terms, and only those specific to the positive class are extended contributing to rebalance positive training samples by reducing the dominance of common terms. It is to be particularly noted that the recall values peak for α in the range of 0.2-0.3, indicating that most severe flu specific terms are selected when the α parameter value is between 0.2 and 0.3. It is also observed that recall value decrease slightly when α drops below 0.2, which is quite expected since specific terms are very rare and it is unlikely to find terms whose frequency in the positive class is at least 5 times higher than their frequency in the negative class. Except for α equal to 0, the system selects very discriminative terms which exist only in the positive class.

As shown in Fig. 3, the precision varies between 55% and a maximum value of 94.25%. It is minimal when α is equal to 1, then evolves by reaching high values when α is in the interval (0.2-0.3) where the majority of specific words are selected. The precision is relatively high because the number of FP is very low which indicates that the system classifies into the positive category only test documents containing specific terms with very high discriminative power.

The weight of the terms has a significant impact on the model performance. The maximum values of recall and precision are reached when the share of the semantic weight in the final weight is around 60%.When the final weight only includes the chi-square weight, i.e. β= 0, the recall and the precision are respectively 72.45% and 83.48%. On the other hand, a final weight comprising only the semantic weight (β= 1), the recall and the precision take the values 78,35% and 88.62% respectively, which means that the semantic weight contributes more to the discriminative power of the terms.

Although SVM performs well on the most commonly used datasets with an F-measure that typically exceeds 93% [33], SVM classification performance is significantly reduced when trained on our dataset using only the BOW feature representation and the TF-IDF weighting scheme, with an F-measure that barely reaches 53%. Which indicates that traditional feature engineering methods are not effective to discriminate between classes sharing many common features.

The results presented in Table III show that our health surveillance system based on a text classification model constructed through an extension of specific features using ELMo and a weighting scheme combining the semantic and chi-square weights is much better than baseline methods.

Experimental results also show that neural word embeddings models (ELMo and word2vec) are more effective than the bag of words and ontology-based approaches.

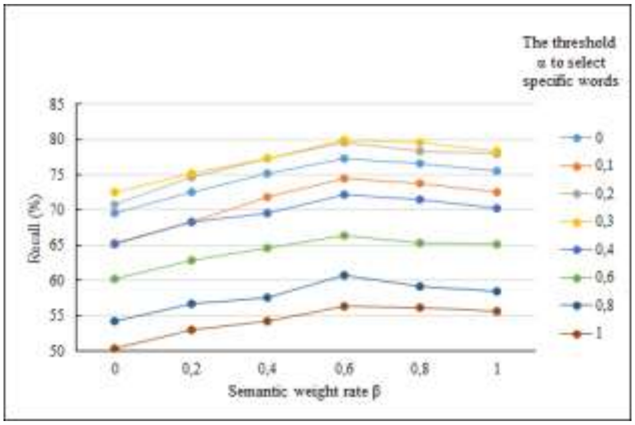

Fig. 2. SVM Classification Recall when Varying Thresholds α and β.

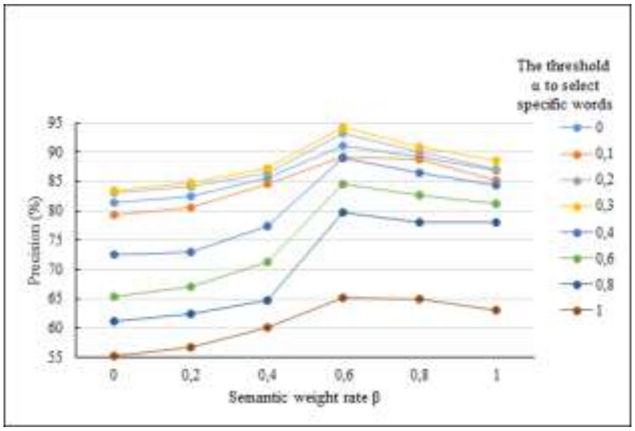

Fig. 3. SVM Classification Precision when Varying Thresholds α and β.

TABLE III. SVM CLASSIFICATION PERFORMANCE OF OUR FEATURE ENGINEERING MODEL COMPARED TO BASELINE METHODS

Features	Recall	Precision	F-measure
Our feature enginnering model	80.00%	94.25%	86.54%
BOW + TF-IDF	52.32%	55.17%	53.70%
UMLS concepts + TF-IDF	65,50%	70,63%	67,96%
Pretrained word2vec	75,23%	80,75%	77.89%

V. CONCLUSION

A system for detecting the occurrence of severe forms of flu by using only clinical texts recorded in EHRs is devised through a text classification model with the challenge of discriminating between severe and mild flu-related documents containing many common features.

To improve classification performance, we have adopted a two-phase approach. In the first phase, with the aim of emphasizing severe flu specific terms deemed rare and discriminative, we have extended these terms by using the pre-trained word embedding ELMo. In the second phase, a combination of two weights is assigned to each term, a semantic weight representing the term's similarity to the word "Pneumonia", and a chi-square weight measuring the correlation between the term and the class "severe flu".

We have found through our experiments that the proposed feature engineering model based on terms extension using deep word representation combined with a weighting scheme that emphasizes discriminative words vigorously improves classification performance when the target classes are very similar.

In this paper, only medical terms are used to determine cases of severe flu. However, opinion words are also useful in deciding the severity of an illness. Thus, mining opinion words occurring in clinical texts is an interesting line of research for our next work.

REFERENCES

[1] P. B. Jensen, L. J. Jensen, et S. Brunak, « Mining electronic health records: towards better research applications and clinical care », Nat. Rev. Genet., vol. 13, no 6, Art. no 6, juin 2012.

[2] S. R. Raman et al., « Leveraging electronic health records for clinical research », Am. Heart J., vol. 202, p. 13-19, 2018.

[3] P. Raghavan, J. L. Chen, E. Fosler-Lussier, et A. M. Lai, « How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? », AMIA Summits Transl. Sci. Proc., vol. 2014, p. 218, 2014.

[4] B. Agarwal et N. Mittal, « Text classification using machine learning methods-a survey », in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 2014, p. 701-709.

[5] L. Yu et H. Liu, « Feature selection for high-dimensional data: A fast correlation-based filter solution », in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, p. 856-863.

[6] D. Wang, H. Zhang, R. Liu, et W. Lv, « Feature selection based on term frequency and T-test for text categorization », in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, p. 1482-1486.

[7] Y. Xu, B. Wang, J. Li, et H. Jing, « An extended document frequency metric for feature selection in text categorization », in Asia Information Retrieval Symposium, 2008, p. 71-82.

[8] P. Samant et R. Agarwal, « Machine learning techniques for medical diagnosis of diabetes using iris images », Comput. Methods Programs Biomed., vol. 157, p. 121-128, 2018.

[9] A. Janecek, W. Gansterer, M. Demel, et G. Ecker, « On the relationship between feature selection and classification accuracy », in New challenges for feature selection in data mining and knowledge discovery, 2008, p. 90-105.

[10] V. Lampos, B. Zou, et I. J. Cox, « Enhancing feature selection using word embeddings: The case of flu surveillance », in Proceedings of the 26th International Conference on World Wide Web, 2017, p. 695-704.

[11] V. N. Garla et C. Brandt, « Ontology-guided feature engineering for clinical text classification », J. Biomed. Inform., vol. 45, no 5, p. 992-998, 2012.

[12] K. Buchan, M. Filannino, et Ö. Uzuner, « Automatic prediction of coronary artery disease from clinical narratives », J. Biomed. Inform., vol. 72, p. 23-32, 2017.

[13] X. Yan et J. Bien, « Rare feature selection in high dimensions », J. Am. Stat. Assoc., p. 1-14, 2020.

[14] K. Patel, D. Patel, M. Golakiya, P. Bhattacharyya, et N. Birari, « Adapting pre-trained word embeddings for use in medical coding », in BioNLP 2017, 2017, p. 302-306.

[15] A. Miaschi et F. Dell'Orletta, « Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation », in Proceedings of the 5th Workshop on Representation Learning for NLP, Online, juill. 2020, p. 110-119.

[16] Ö. Uzuner, B. R. South, S. Shen, et S. L. DuVall, « 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text », J. Am. Med. Inform. Assoc., vol. 18, no 5, p. 552-556, 2011.

[17] G. Shmueli et H. Burkom, « Statistical challenges facing early outbreak detection in biosurveillance », Technometrics, vol. 52, no 1, p. 39-51, 2010.

[18] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, et R. A. Dudley, « N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit », J. Am. Med. Inform. Assoc., vol. 21, no 5, p. 871-875, 2014.

[19] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, et N. Grayson, « Automatic ICD-10 classification of cancers from free-text death certificates », Int. J. Med. Inf., vol. 84, no 11, p. 956-965, 2015.

[20] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, et M. Yetisgen-Yildiz, « Pneumonia identification using statistical feature selection », J. Am. Med. Inform. Assoc., vol. 19, no 5, p. 817-823, 2012.

[21] M. Conway, J. Dowling, et W. Chapman, « Developing an application ontology for mining free text clinical reports: the Extended Syndromic Surveillance Ontology », in Proceedings of the Third International Workshop on Health Document Text Mining and Information Analysis, Slovenia (LOUHI 2011), 2011, p. 75-82.

[22] M. Alobaidi, K. M. Malik, et M. Hussain, « Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain », Comput. Methods Programs Biomed., vol. 165, p. 117-128, 2018.

[23] Y. Zhang, Q. Chen, Z. Yang, H. Lin, et Z. Lu, « BioWordVec, improving biomedical word embeddings with subword information and MeSH », Sci. Data, vol. 6, no 1, p. 1-9, 2019.

[24] Y. Fan, S. Pakhomov, R. McEwan, W. Zhao, E. Lindemann, et R. Zhang, « Using word embeddings to expand terminology of dietary supplements on clinical notes », JAMIA Open, vol. 2, no 2, p. 246-253, 2019.

[25] W. Zhu, W. Zhang, G.-Z. Li, C. He, et L. Zhang, « A study of damp-heat syndrome classification using Word2vec and TF-IDF », in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, p. 1415-1420.

[26] X. Luo et S. Shah, « Concept embedding-based weighting scheme for biomedical text clustering and visualization », in Applied Informatics, 2018, vol. 5, no 1, p. 1-19.

[27] A. R. Aronson, « Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. », in Proceedings of the AMIA Symposium, 2001, p. 17.

[28] T. Mikolov, K. Chen, G. Corrado, et J. Dean, « Efficient estimation of word representations in vector space », ArXiv Prepr. ArXiv13013781, 2013.

[29] J. Pennington, R. Socher, et C. D. Manning, « Glove: Global vectors for word representation », in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, p. 1532-1543.

[30] M. E. Peters et al., « Deep contextualized word representations », ArXiv Prepr. ArXiv180205365, 2018.

[31] Q. Jin, B. Dhingra, W. W. Cohen, et X. Lu, « Probing Biomedical Embeddings from Language Models », ArXiv190402181 Cs, avr. 2019, Consulté le: janv. 28, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1904.02181.

[32] R. Karsi, M. Zaim, et J. El Alami, « Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques », in Journal of Physics: Conference Series, 2017, vol. 870, no 1, p. 012005.

[33] Z. Liu, X. Lv, K. Liu, et S. Shi, « Study on SVM compared with the other text classification methods », in 2010 Second international workshop on education technology and computer science, 2010, vol. 1, p. 219-222.