IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 4, August, 2021, pp. 307-314
Journal homepage: http://iieta.org/journals/ria

Leveraging Pre-Trained Contextualized Word Embeddings to Enhance Sentiment Classification of Drug Reviews

Redouane Karsi*, Mounia Zaim, Jamila El Alami

LASTIMI Laboratory, Higher School of Technology of Sale, Mohammed V University, Rabat 10000, Morocco

Corresponding Author Email: rdkarsi@yahoo.fr

https://doi.org/10.18280/ria.350405

Received: 21 June 2021
Accepted: 13 August 2021

Keywords:
contextual word embedding, drug reviews, ELMo, machine learning, pre-trained word embedding, sentiment analysis

ABSTRACT

Traditionally, pharmacovigilance data are collected during clinical trials on a small sample of patients and are therefore insufficient to adequately assess drugs. Nowadays, consumers use online drug forums to share their opinions and experiences about medication. These feedbacks, which are widely available on the web, are automatically analyzed to extract relevant information for decision-making. Currently, sentiment analysis methods are being put forward to leverage consumers' opinions and produce useful drug monitoring indicators. However, these methods' effectiveness depends on the quality of word representation, which presents a real challenge because the information contained in user reviews is noisy and very subjective. Over time, several sentiment classification problems use machine learning methods based on the traditional bag of words model, sometimes enhanced with lexical resources. In recent years, word embedding models have significantly improved classification performance due to their ability to capture words' syntactic and semantic properties. Unfortunately, these latter models are weak in sentiment classification tasks because they are unable to encode sentiment information in the word representation. Indeed, two words with opposite polarities can have close word embeddings as they appear together in the same context. To overcome this drawback, some studies have proposed refining pre-trained word embeddings with lexical resources or learning word embeddings using training data. However, these models depend on external resources and are complex to implement. This work proposes a deep contextual word embeddings model called ELMo that inherently captures the sentiment information by providing separate vectors for words with opposite polarities. Different variants of our proposed model are compared with a benchmark of pre-trained word embeddings models using SVM classifier trained on Drug Review Dataset. Experimental results show that ELMo embeddings improve classification performance in sentiment analysis tasks on the pharmaceutical domain.

1. INTRODUCTION

Pharmacovigilance is an integral part of the health surveillance system aimed at assessing the effectiveness and side effects of medicines. Drug testing involves the interpretation of data collected during clinical trials on volunteer patients. However, it is difficult to provide reliable results as data is obtained from a small sample of volunteers over a limited period [1].

The emergence of pharmaceutical-related platforms such as blogs and discussion forums has paved the way for people to share on the web their opinions and experiences with the medicines they take [2]. These consumer opinions feed pharmacovigilance systems with clues to help monitor drugs. However, unlike data obtained from clinical trials, the data collected from the internet is unstructured, and we need to use natural language processing (NLP) techniques to leverage this data. Sentiment analysis is the most appropriate NLP task for analyzing the subjective information in consumer reviews to identify the opinion orientation of the text.

In sentiment analysis, two approaches can be distinguished. The lexical-based approach [3] uses a dictionary to assign a sentiment value to each word in the document. The machine learning-based approach [4] utilizes a classification algorithm trained on text corpora. Previous research has revealed that machine learning methods deliver the best performance [5].

Text representation is a key element that affects the performance of machine learning algorithms. The straightforward bag of words model (BOW) is by far the most widely used to represent textual data. However, this model is not perfect because BOW vectors are sparse and have a high dimensionality equivalent to the vocabulary size, ignoring the semantic links between words [6]. In recent years, word embeddings models learned from neural networks like Word2vec [7] and GloVe [8] have proven to improve text representation by capturing into a dense vector the syntactic and semantic information of each word in the vocabulary. As a result, words appearing in the same contexts have neighboring vectors.

Unfortunately, conventional word embeddings models fail to capture sentiment information conveyed by opinion features [9], which means that words with similar vector representation may have opposite sentiment orientation, thus reducing sentiment classification performance. For instance, consider the following review:

The vaccine is effective, but I experienced a severe allergic reaction.

In the example above, the words *"effective"* and *"severe"*

express opposite sentiment polarity, yet they occur together in drug reviews and are assigned similar vectors.

Thus, word embedding models are adjusted to take into account the sentiment information ignored by traditional models. In this sense, several studies have proposed models to simultaneously incorporate syntactic, semantic, and sentiment properties into the vector representation of words using two main approaches. The first approach leverages labeled data to learn embeddings using a supervised neural network [10]. The second approach is applied to the pre-trained vectors obtained from word embeddings models. The idea being to use a sentiment dictionary to refine the pre-trained vector of each opinionated word so that it is closer to neighboring words with the same sentiment and semantic orientation and further away from those with an opposite sentiment polarity [11].

However, both approaches suffer from some limitations that reduce their performance. Indeed, training word embeddings in a supervised manner requires a large amount of labelled data which is not sufficiently available in the medical domain [12]. In addition, refinement methods use a general lexicon that cannot capture the exact meaning of domain-dependent words [13]. For instance, in the sentence *"I have tested positive for COVID-19"*, the word *"positive"* expresses a negative sentiment, whereas it is labeled positive in the general lexicon. Moreover, sentiment dictionaries do not cover the entire corpus vocabulary and do not provide entries for informal words and abbreviations commonly used to express sentiments in the medical domain [14].

To overcome these limitations, contextual word embeddings such as ELMo [15] and BERT [16] have proven to be very effective in managing out of vocabulary and polysemous words. These models produce word vectors that vary depending on the surrounding context and may better capture affective words' sentiment orientation [17]. In this paper, we propose different pre-trained ELMo word embeddings to train the SVM classifier in sentiment analysis task on drug reviews dataset. Results show that combining ELMo embeddings pre-trained respectively on general and medical domain outperform baseline word representations. Our main contributions in this work are as follows:

- We performed various experiments to compare the performance of different embeddings in the sentiment analysis task on drug reviews. We have shown that ELMo representations outperform traditional word embeddings. Furthermore, we found that the concatenation of ELMo embeddings pre-trained respectively on the general and medical domains yield the best performance.
- We show that dimensionality reduction using the principal component analysis (PCA) [18] method slightly affects classification performance while substantially lowering training time.
- We demonstrate that the proposed ELMo models, and more precisely the concatenation of the general and medical domains produces word representations that efficiently learn the SVM classifier with short texts and little training data, proving that these models are adapted to the drug domain in the sentiment analysis task.

This paper is structured as follows. In part 2, some previous work on word representations for the sentiment analysis task are reviewed. Our methodology is described in part 3. In part 4, we present and discuss the experimental results. Finally, conclusions are given in part 5.

2. RELATED WORK

For many years, sentiment analysis has been used broadly in many domains such as movies, politics, and hotels. However, little interest has been shown in the pharmaceutical area. In recent studies, researchers have tackled text classification tasks in drug reviews by exploring lexicon [19], machine learning [20], and role-based approaches [21].

In the work [22], the authors experimented with different bag-of-words representations (unigrams and ngrams) to analyse patients' online comments about the quality of health care. They used four machine learning algorithms, namely Naïve Bayes multinomial, Bagging, Decision trees, and SVM. They found that all these algorithms achieve reasonable accuracies. The authors in Ref. [23] performed six machine learning algorithms to determine the sentimental orientation of tweets about side effects. Results have shown that SVM trained with unigram, bigram, and WordNet gives the best performance. The Cross-domain sentiment analysis approach is proposed by Gräßeret et al. [24] to determine the polarity of different aspects (Overall Rating, Effectiveness and side effects) in drug reviews. They concluded that performance depends on the training domain. A method based on syntactic dependency paths to extract and classify aspects from drug reviews is proposed in Ref. [25]. The solution shows interesting results compared to baseline methods. To remedy the poor performance of sentiment classification due to noisy data in online reviews, the work described by Asghar et al. [26] provides an approach using rule-based classification scheme by handling domain specific words, negations, emoticons, and modifiers, which help to improve all performance indicators.

A word embedding model maps a word to a real vector that captures syntactic and semantic information so that similar words have neighboring vectors. Carrillo-de-Albornoz et al. [27] trained an SVM classifier with a manually labeled dataset constructed from health-related forums concerning breast cancer to classify the sentences of each post into three different categories: Opinion, fact, or experience. Testing with several types of features, they found that word representations based on bag-of-words and word embedding models provide the best results. In contrast, domain-specific features such as semantic types and UMLS concepts produce modest results. The work [28] confirms through different machine learning techniques that the combination of sentiment features, lexical features, and word embeddings achieves high accuracy on a dataset containing patient-authored data.

The works presented above show the superiority of word embeddings over the bag of words model in sentiment classification. However, these models are not optimal as they fail to capture sentiment information that may result in words with opposite polarities having similar vectors. Thus, two approaches have emerged to encode sentiment orientation in the word representation.

Maas et al. [29] provided a model that combines an unsupervised probabilistic component to learn semantic and syntactic properties of words and another component that takes advantage of a large annotated dataset to integrate sentiment information into the word representation. Sentiment-specific word embedding architecture is described in Ref. [30]. It uses a global representation based on a corruption strategy in conjunction with a local representation to capture sentiment and semantic information from words, respectively. In the study [31], the authors exploited the SWN lexicon to extract medical sentiment features and use them as

input to train different machine learning algorithms using word embeddings. The results indicate that the Med-SWN lexicon outperforms all other types of word embeddings. The method suggested by Ye et al. [32] employs an external lexical sentiment combined with supervised training data to fine-tune pre-trained word vectors using a CNN sentiment classifier. The method achieves state-of-the-art performance over four benchmark sentiment analysis datasets. In the study detailed in Ref. [11]. The authors proposed a model that can improve both sentiment embeddings and conventional word embeddings. It is based on a sentiment intensity lexicon for refining pre-trained word vectors so that they are close to their sentimentally similar neighbors and not far from their original vectors. A method called RGWE is proposed by Wang et al. [33]. It aims to refine the word vectors generated by GloVe and Word2Vec by embedding different features like POS and sentiment concepts. Their method uses a multi semantics sentiment intensity lexicon to provide optimal sentiment information that can better represent polysemous words.

The two approaches mentioned above are penalized by the lack of labeled data and the low coverage of the sentiment lexicon, especially in the medical domain. Besides, they do not handle polysemous words efficiently.

In recent research, contextual word embeddings are adopted to remedy the shortcomings of static models in capturing an accurate representation of words, especially contextual information such as sentiment.

In a comparative study [34], contextual word embeddings outperformed traditional word embeddings on clinical concept extraction tasks. Moreover, embeddings pre-trained on a clinical dataset achieve much better performance than those pre-trained on an open domain corpus. The study [35] states that the clinical named entity recognition can be improved by combining ELMo and Flair methods pre-trained on clinical data. The work reported in Ref. [36] shows that different contextual word embeddings give the best results in three classification issues in the health domain.

We remark that previous work has not sufficiently explored contextual embeddings to address sentiment analysis of drug reviews. In this work, we demonstrate through a comparative study that contextual embeddings improve sentiment classification performance in the drug domain.

3. METHODOLOGY

Our proposed sentiment classification model consists of two main steps. In the first step, after preprocessing all reviews, we build word vectors using different pre-trained word embeddings. Then, each review is represented in the vector space by averaging the embeddings of its constituent words to train an SVM classifier in order to predict the sentiment polarity of drug reviews. An overview of our sentiment classification process is depicted in Figure 1.

Our baseline work presented by Si et al. [34] proposed different pretrained word embeddings representations to enhance clinical concept extraction task. The authors aim to compare the performance of different word representations in four clinical concept extraction tasks. The methodology followed consists in using several state-of-the-art word embedding techniques, namely: Word2vec, GloVe, fastText, ELMo, BERT pre-trained in both the general and medical domains to generate word vectors. Then, the generated embeddings are fed as input to a biLSTM neural network. The bidirectional LSTM is made up of two LSTM networks: one that accepts input in a forward direction and the other that takes input in the backward direction. When the outputs of the two networks are combined, the context surrounding each individual word is captured. The biLSTM output is then fed into a linear CRF to make predictions based on this improved data. The results show significant improvements when word embeddings models are pretrained on clinical domain.

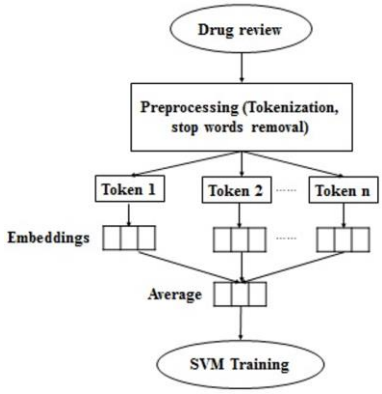

Figure 1. Our sentiment classification process

Our approach aims at extending the models proposed in our baseline work in order to achieve better performance in sentiment analysis task on the pharmaceutical domain. Contextual word embeddings have proven their superiority over static methods. Indeed, in our baseline work, the ELMo embeddings pre-trained on the MIMIC dataset offers the best results in clinical concept extraction tasks. Thus, our proposed word representation relies mainly on ELMo as it better captures sentiment information and correctly represents out-of-vocabulary words. The novelty of our approach is to experiment with other variants of ELMo embeddings and their combinations to improve sentiment classification performance on drug reviews.

3.1 Deep Contextualized Word Representations (ELMo)

ELMo is a recent model of word representation introduced by AllenNLP [15]. The ELMo model produces different embeddings for the same word depending on the context where it occurs. Unlike traditional models that use lookup tables to generate word embeddings, ELMo relies on a pre-trained language model. Formally, the goal of language model is to assign a probability to every term in a sequence of words such as sentences.

ELMo is based on a language model called biLSTM, which combines two LSTMs that process word sequences in both forward and backward directions. For each word, the internal states calculated from both the forward and backward passes are concatenated to produce an intermediate word vector. ELMo uses multiple biLSTM layers that are stacked together so that the intermediate word vector produced by the bottom layer is fed into the next higher layer. This process is repeated until the last layer is reached. Thus, in the upper layers, the internal states capture the more abstract semantic features such as sentiment. The final vector provided as input to downstream tasks is a weighted combination of the intermediate vectors produced by each biLSTM layer and the input vector in a task-dependent manner.

In text classification with CNN, sentences are treated at

word level. Thus, misspelled and out-of-vocabulary words are not correctly represented. To overcome this problem, ELMo uses the character-based CNN approach which processes sentences at the character level so that unknown words are reduced to extract meaningful features improving classification performance. Thus, ELMo feeds the first biLSTM layer with word embeddings computed only from characters using a character-based CNN helping to handle out-of-vocabulary and misspelled words. For example, "sickkk" and "sick" have similar vectors. The ELMo model architecture is shown in Figure 2.

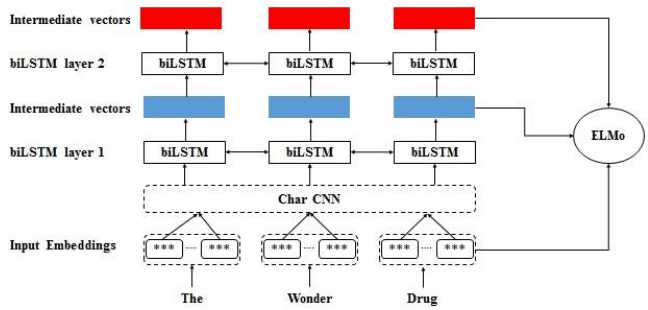

Figure 2. The ELMo model architecture

3.2 Static word embeddings

We employed two static word embeddings to train the SVM classifier to predict sentiment polarity of drug reviews. These word representations are detailed below.

- **Word2vec**: Word2vec is a word embeddings algorithm. It was developed by a Google research team. It is based on two-layer neural networks and seeks to learn word vectors, so that words that share similar contexts are represented by close vectors. Word2vec has two neural architectures, called CBOW and Skip-Gram. CBOW receives as input the context of a target word, i.e. the terms surrounding it in a sentence, and tries to predict the target word. In contrast, Skip-Gram takes as input a word and tries to predict its context. In both cases, the network is trained by going through the provided text and modifying the neural weights to reduce the prediction error of the algorithm.
- **GloVe:** Unlike Word2vec that depends on the local context of words, GloVe is an unsupervised algorithm that employs global statistics on word co-occurrence to create word embeddings. It combines global matrix factorisation and local context window method. The model constructs a large matrix of co-occurrence where the rows represent words and the columns are contexts. Each matrix element corresponds to the frequency of a word in a given context defined by a window-size. Then the matrix is factorized by minimizing a reconstruction loss to yield a lower-dimensional representation, so that each row represents the vector representation of each word.

3.3 Pre-trained word embeddings

In our study, we used the following pre-trained models:

- **ELMo**, **Word2vec** and **GloVe** as trained in [34]. The authors used the MIMIC III dataset which is composed of about 2 million clinical notes collected from the Intensive Care Unit. Word2Vec and Glove generate 300-dimension embeddings, whereas ELMo generates 1024-dimension embeddings.
- The **original ELMo** model [15] is trained on a dataset containing a mixture of Wikipedia and online news consisting of 5.5B tokens. It generates 1024-dimensional embeddings.
- **BioELMo** is a model produced using the Tensorflow implementation of ELMo [37]. It is trained on 10 million recent abstracts of medical articles collected from pubMed providing a word representation adapted to the biomedical domain. BioELMo embeddings are of dimension 1024.

3.4 The dataset

The SVM classifier is trained on a dataset of drug reviews [24] collected from the drugs.com website. The users' feedback towards a drug is expressed as a score between 1 and 10. The dataset is composed of about 200,000 reviews differently distributed according to the scores reported by the users, with a clear predominance of reviews with the scores (1, 2, 9 and 10).

After sorting reviews by user rating, we manually and randomly extracted a 100k balanced dataset between positive and negative samples so that reviews with scores in the interval (1,4) are assigned the negative class, while reviews with scores above 7 are assigned the positive class. We consider reviews with scores of 5 and 6 conveying neutral sentiment and are not included in our dataset. Moreover, the different sentiment intensities are represented in our dataset with the distribution detailed in Table 1. At last, we split the data between the training and test sets in an 80-20 ratio.

Table 1. Reviews distribution on the whole and the 100k datasets according to the sentiment intensity

Sentiment intensity	Number of reviews (Whole dataset)	Number of reviews (100k dataset)
1	28918	25346
2	9265	9265
3	8718	8718
4	6671	6671
7	12547	12500
8	25046	12500
9	36708	12500
10	68005	12500

3.5 Word embeddings generation

Before generating word embeddings, the textual reviews undergo pre-processing operations including tokenisation and stop words removal such as punctuation and articles. These actions serve to clean up the text data and therefore avoid producing noisy word embeddings that degrade classification performance.

ELMo is a language model that predicts the next word given a sequence of words such as sentences. To generate embeddings, ELMo receives as input sequences of the same length. Thus, in our case, we set the size of the reviews to 100 tokens since most comments are 100 tokens or less. To do this, we truncate or pad the text to equal the maximum length.

The objective of this work is to compare the performance of different pre-trained embeddings when used to generate inputs for training an SVM classifier in the sentiment classification task on the pharmaceutical domain. We use as a baseline the embeddings experimented in Ref. [34] on the clinical concept extraction task namely Word2vec, GloVe and ELMo pre-trained in the MIMIC dataset. Then, we enrich our comparative study with two ELMo variants: BioELMo and the

original ELMo supposed to capture better sentiment information if combined together. More details on the pre-trained embeddings used in our study are presented in Table 2.

Table 2. Details of the pre-trained embeddings used in our study

Embeddings	Corpus	Dimension
Word2vec	MIMIC III	300
GloVe	MIMIC III	300
ELMo	MIMIC III	1024
Original ELMo	Wikipedia + WMT 2008-2012	1024
BioELMo	PubMed	1024

4. EXPERIMENTAL RESULTS AND DISCUSSION

In our comparative study, we specifically analysed the impact of three parameters on the performance of different word embeddings: word embeddings dimensionality, dataset size, and word sequence length. We use the F-score to evaluate the classification performance, which is the harmonic mean between precision and recall. These performance indicators are calculated as follows:

$$Precision = \frac{T_P}{T_P + F_P} \quad (1)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (2)$$

$$F - score = 2.\frac{Precision.Recall}{Precision + Recall} \quad (3)$$

where,

- T_P (True Positive): correctly predicted as positive.
- F_P (False Positive): incorrectly predicted as positive.
- T_N (True Negative): correctly predicted as negative.
- F_N (False Negative): incorrectly predicted as negative.

4.1 Impact of embedding dimension

The pre-trained word embeddings are of different dimensions, and the vectors resulting from the concatenation of two embeddings are high dimensional and cause a long training time. To mitigate this failure, many techniques are used to lower the dimensionality of the vectors while keeping the most important data. Among these techniques, we use PCA to evaluate the effect of embeddings dimensionality on sentiment classification performance.

Referring to Figure 3, we remark that the original vector of all embeddings performs best with a maximum value of 88.2% achieved by Original ELMo and BioELMo concatenation and a minimum value of 71.36% achieved by Word2vec (MIMIC). Moreover, Word2vec (MIMIC), GloVe (MIMIC), and ELMo (MIMIC) perform less well in the sentiment classification task compared to the baseline performance recorded in the clinical concept extraction task. On the other hand, we observe that the reduction of dimensionality leads to a decrease in performance. Thus, when the dimensionality of the vectors is reduced and remains above 400, the performance decreases by between 0.59% and 1.89% percentage points. This drop in performance becomes more significant with a loss of between 7.51% and 12.4% percentage points for a dimensionality below 400. Regarding the time needed to complete the SVM classifier training, we observed the best performing word embeddings behavior, i.e., Original ELMo + BioELMo. Hence, we can see through Figure 4 that reducing the original vector from 2048 to 400 dimensions, we save 45% of the training time while ensuring relatively stable performances between 87.61% and 88.2%.

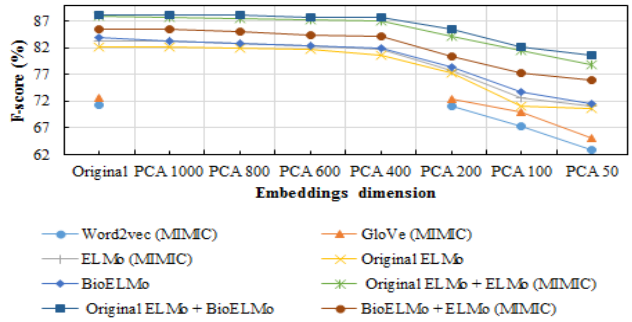

Figure 3. Performance of different word representations when varying the embedding dimension

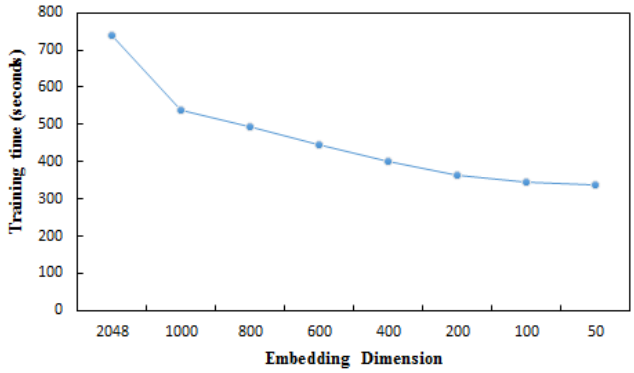

Figure 4. Training time of Original ELMo + BioELMo when varying the embedding dimension

4.2 Impact of dataset size

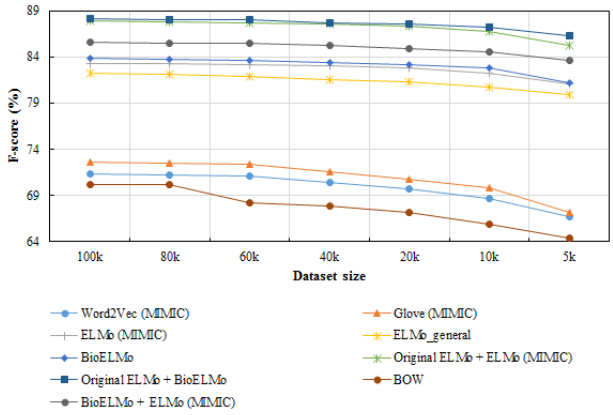

Figure 5. Performance of different word representations when varying the dataset size

In our second experiment, we train the SVM classifier with different sized datasets extracted from the original dataset comprised of 100k reviews. As shown in Figure 5, the performance of different embeddings when reducing the dataset size indicates that the neural network based word embeddings significantly outperform the traditional BOW

model, and contextual word embeddings outperform traditional embeddings regardless of the dataset size.

However, once the dataset size drops below 20k, the performance decreases at a steeper rate reaching a minimum overall average F-score of 67.68% with a standard deviation of 2.13 for the bag-of-words model, then a maximum overall average F-score of 87.58% with a standard deviation of 0.65 for the Original ELMo + BioELMo concatenation embeddings.

4.3 Impact of sequence length

In this experiment, drug reviews are represented as sequences of words with different sizes ranging from 20 to 100 tokens. The results presented in Figure 6 show that the performance of the bag-of-words model and traditional embeddings deteriorates significantly when the sequence length is below 50, with respective decreases of around 4 and 5 percentage points. On the other hand, contextual embeddings do better by keeping a good performance even if the sequence length is only 30 words. Indeed, ELMo (MIMIC), Original ELMo, and BioELMo consistently achieve F-scores above 74%. Meanwhile, the concatenations Original ELMo + BioELMo and Original ELMo + ELMo (MIMIC) and BioELMo + ELMo (MIMIC) outperform all other representations with an F-score above 79% and a loss of almost 2 percentage points at the most. Furthermore, unlike the two previous experiments, we see that Original ELMo does better than ELMo (MIMIC) as the sequence length decreases.

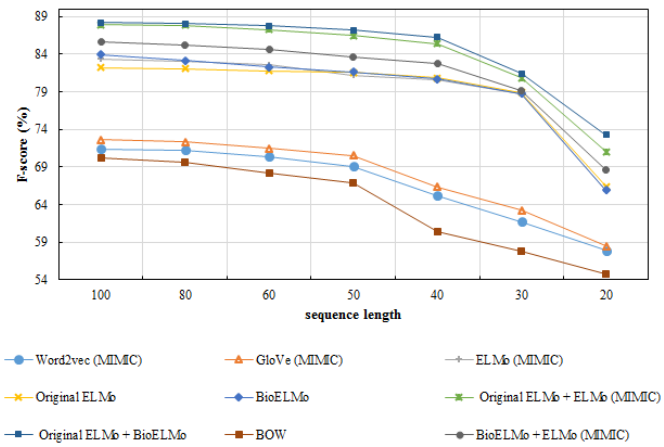

Figure 6. Performance of different word representations when varying the sequence size

4.4 Combining threshold values of the three parameters

The different experiments we have conducted confirm that ELMo embeddings concatenation provides the best performance. We identified threshold values for the three parameters: (dimension size = 400, dataset size = 20k, sequence length = 30), above which the performance remains stable and close to its maximum values. We evaluated all ELMo embeddings using threshold values of the three parameters. As shown in Table 3, we found that ELMo performs better than traditional embeddings evaluated with the original parameter values. Furthermore, the concatenation of ELMo embeddings yields an F-score of over 78%, with an apparent decrease in training time from 740 seconds to almost 80 seconds.

Table 3. F-score and training time of different ELMo embeddings using threshold values of parameters (Embeddings dimensionality, dataset size and sequence length)

Embeddings	F-score	Training time (s)
ELMo (MIMIC)	77.20%	82
Original ELMo	77.42%	76
BioELMo	77.32%	74
BioELMo + ELMo (MIMIC)	78.36%	76
Original ELMo + ELMo (MIMIC)	80.03%	81
Original ELMo + BioELMo	80.60%	78

4.5 Discussion

The performance comparison of different word embeddings in sentiment classification tasks on the pharmaceutical domain shows that contextual embeddings perform better than traditional embeddings. This result is due to the fact that sentiment information is context-dependent and therefore better captured by contextual embeddings. Whereas traditional embeddings assign similar vectors to words that appear together even if they have opposite polarity, thereby failing to capture sentiment information.

About the impact of embedding dimensionality, we find that the PCA method solves the problem of high dimensionality affecting contextual embeddings. Hence, despite a decrease in performance due to the reduction in dimensionality leading to a loss of information contained in word embeddings, the PCA method proves its effectiveness since it manages to maintain an F-score higher than 80% even when the dimension is reduced from 2048 to 400, contributing at the same time to reduce the training time.

Concerning the impact of the dataset size, unlike the bag-of-words model, ELMo embeddings and traditional embeddings are insensitive to the decrease of the dataset size and keep relatively stable performances. Indeed, the bag-of-words model needs many training samples to derive semantic similarities between words. In contrast, word embeddings are already pre-trained on large corpora and require little training data to learn the classifier. This last result shows that by using pretrained word embeddings, sentiment classification tasks can be tackled with little labelled data. However, when the dataset size is below 20k the performance decreases faster. This is because the amount of training data is not sufficient for the SVM classifier to provide good predictions.

Regarding the impact of word sequence length, the results show that ELMo embeddings manage to capture sentiment information in a short context, proving that ELMo is a powerful model pre-trained on large corpora suitable for short text classification in the pharmaceutical domain. Furthermore, we note that ELMo pre-trained on the general domain is better than its variant pre-trained on the medical domain in short text classification tasks, probably because the first sentences of reviews use general vocabulary.

Finally, among all the ELMo variants we tested, the combination of Original ELMo and BioELMo provides the best performance, even surpassing the combination of the two medical domains: ELMo (MIMIC) and BioELMo, because it is based on a vast vocabulary covering both the general and medical domains, thus approximating the way people express their opinions about drugs.

5. CONCLUSIONS

In this work, we conducted the sentiment analysis task on drug reviews by addressing the problem faced by traditional word representation models to encode sentiment information. We evaluated ELMo embeddings pre-trained on both medical and general domains compared to previously experimented embeddings in clinical concept extraction tasks. We found that ELMo embeddings outperform traditional embeddings. The best performance comes from the concatenation of the general and medical domains. We also show that the proposed ELMo models are little affected by dimensionality reduction. Moreover, they effectively classify short reviews and are best adapted to small training data.

REFERENCES

[1] Luo, Y., Thompson, W.K., Herr, T.M., Zeng, Z., Berendsen, M.A., Jonnalagadda, S.R. (2017). Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Safety, 40(11): 1075-1089. https://doi.org/10.1007/s40264-017-0558-6

[2] Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S. (2015). Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics, 54: 202-212. https://doi.org/10.1016/j.jbi.2015.02.004

[3] Khoo, C.S., Johnkhan, S.B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science, 44(4): 491-511. https://doi.org/10.1177/0165551517703514

[4] Aydoğan, E., Akcayol, M.A. (2016). A comprehensive survey for sentiment analysis tasks using machine learning techniques. In 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), pp. 1-7. https://doi.org/10.1109/INISTA.2016.7571856

[5] Zhang, H., Gan, W., Jiang, B. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. In 2014 11th Web Information System and Application Conference, pp. 262-265. https://doi.org/10.1109/WISA.2014.55

[6] Zhao, R., Mao, K. (2017). Fuzzy bag-of-words model for document representation. IEEE Transactions on Fuzzy Systems, 26(2): 794-804. https://doi.org/10.1109/TFUZZ.2017.2690222

[7] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[8] Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

[9] Yu, L.C., Wang, J., Lai, K.R., Zhang, X. (2017). Refining word embeddings for sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 534-539. https://doi.org/10.1109/TASLP.2017.2788182

[10] Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., Wang, Q. (2017). Weakly-supervised deep embedding for product review sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 30(1): 185-197. https://doi.org/10.1109/TKDE.2017.2756658

[11] Yu, L.C., Wang, J., Lai, K.R., Zhang, X. (2017). Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(3): 671-681. https://doi.org/10.1109/TASLP.2017.2788182

[12] Guan, Z., Chen, L., Zhao, W., Zheng, Y., Tan, S., Cai, D. (2016). Weakly-Supervised Deep Learning for Customer Review Sentiment Classification. In IJCAI, pp. 3719-3725.

[13] Muhammad, A., Wiratunga, N., Lothian, R., Glassey, R. (2013). Domain-Based lexicon enhancement for sentiment analysis. In SMA@ BCS-SGAI, pp. 7-18.

[14] Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C. (2017). Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications, 69: 214-224. https://doi.org/10.1016/j.eswa.2016.10.043

[15] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[16] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[17] Ilić, S., Marrese-Taylor, E., Balazs, J.A., Matsuo, Y. (2018). Deep contextualized word representations for detecting sarcasm and irony. arXiv preprint arXiv:1809.09795.

[18] Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.

[19] Liu, S., Lee, I. (2019). Extracting features with medical sentiment lexicon and position encoding for drug reviews. Health Information Science and Systems, 7(1): 1-10. https://doi.org/10.1007/s13755-019-0072-6

[20] Daniulaityte, R., Chen, L., Lamy, F.R., Carlson, R.G., Thirunarayan, K., Sheth, A. (2016). "When 'bad' is 'good'": Identifying personal communication and sentiment in drug-related tweets. JMIR Public Health and Surveillance, 2(2): e6327. https://doi.org/10.2196/publichealth.6327

[21] Na, J.C., Kyaing, W.Y.M., Khoo, C.S., Foo, S., Chang, Y.K., Theng, Y.L. (2012). Sentiment classification of drug reviews using a rule-based linguistic approach. In International Conference on Asian Digital Libraries, pp. 189-198.

[22] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. Journal of Medical Internet Research, 15(11): e239. https://doi.org/10.2196/jmir.2721

[23] Yu, F., Moh, M., Moh, T.S. (2016). Towards extracting drug-effect relation from Twitter: A supervised learning approach. In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), pp. 339-344. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.53

[24] Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In

Proceedings of the 2018 International Conference on Digital Health, pp. 121-125. https://doi.org/10.1145/3194658.3194677

[25] Cavalcanti, D., Prudêncio, R. (2017). Aspect-based opinion mining in drug reviews. In EPIA Conference on Artificial Intelligence, pp. 815-827. https://doi.org/10.1007/978-3-319-65340-2_66

[26] Asghar, M.Z., Khan, A., Ahmad, S., Qasim, M., Khan, I.A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLOS ONE, 12(2): e0171649. https://doi.org/10.1371/journal.pone.0171649

[27] Carrillo-de-Albornoz, J., Aker, A., Kurtic, E., Plaza, L. (2019). Beyond opinion classification: Extracting facts, opinions and experiences from health forums. PLOS ONE, 14(1): e0209961. https://doi.org/10.1371/journal.pone.0209961

[28] Carrillo-de-Albornoz, J., Rodriguez Vidal, J., Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. PLOS ONE, 13(11): e0207996. https://doi.org/10.1371/journal.pone.0207996

[29] Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142-150.

[30] Fu, P., Lin, Z., Yuan, F., Wang, W., Meng, D. (2018). Learning sentiment-specific word embedding via global sentiment representation. In Proceedings of the AAAI Conference on Artificial Intelligence.

[31] Liu, S., Lee, I. (2018). Sentiment classification with medical word embeddings and sequence representation for drug reviews. In International Conference on Health Information Science, pp. 75-86. https://doi.org/10.1007/978-3-030-01078-2_7

[32] Ye, Z., Li, F., Baldwin, T. (2018). Encoding sentiment information into word vectors for sentiment analysis. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 997-1007.

[33] Wang, Y., Huang, G., Li, J., Li, H., Zhou, Y., Jiang, H. (2021). Refined global word embeddings based on sentiment concept for sentiment analysis. IEEE Access, 9: 37075-37085. https://doi.org/10.1109/ACCESS.2021.3062654

[34] Si, Y., Wang, J., Xu, H., Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association, 26(11): 1297-1304. https://doi.org/10.1093/jamia/ocz096

[35] Jiang, M., Sanger, T., Liu, X. (2019). Combining contextualized embeddings and prior knowledge for clinical named entity recognition: Evaluation study. JMIR Medical Informatics, 7(4): e14850. https://doi.org/10.2196/14850

[36] Joshi, A., Karimi, S., Sparks, R., Paris, C., MacIntyre, C.R. (2019). A comparison of word-based and context-based representations for classification problems in health informatics. Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 135-141. https://doi.org/10.18653/v1/W19-5015

[37] Jin, Q., Dhingra, B., Cohen, W.W., Lu, X. (2019). Probing biomedical embeddings from language models. Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, pp. 82-89. https://doi.org/10.18653/v1/W19-2011