A study on the relevance of generic word embeddings for sentence classification in hepatic surgery

Achir Oukelmoun, Nasredine Semmar, Gaël de Chalendar, Enguerrand Habran, Eric Vibert, Emma Goblet, Mariame Oukelmoun, Marc-Antoine Allard

A Study on the Relevance of Generic Word Embeddings for Sentence Classification in Hepatic Surgery

Achir Oukelmoun*†, Nasredine Semmar*, Gaël de Chalendar*
Enguerrand Habran†, Eric Vibert†, Emma Goblet†, Mariame Oukelmoun‡ and Marc-Antoine Allard†
* Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
† Chaire BOPA, Rue de la Chapelle de l'Hôpital, 94800 Villejuif, France
‡ Centre Hospitalier Cheikh Zaid, 10000 Rabat, Morocco

***Abstract*—While the fine-tuning process of extensive contextual language models often demands substantial computational capacity, utilizing generic pre-trained models in highly specialized domains can yield suboptimal results. This paper aims to explore an innovative approach to derive pertinent word embeddings tailored to a specific domain with limited computational resources (The introduced methodologies are tested within the domain of hepatic surgery, utilizing the French language.). This exploration takes place within a context where computational limitations prohibit the fine-tuning of large language models. A new embedding (referred to as FTW2V) that combines Word2Vec and FastText is introduced. This approach addresses the challenge of incorporating terms absent from Word2Vec's vocabulary. Furthermore, a novel method is used to evaluate the significance of word embeddings within a specialized corpus. This evaluation involves comparing classification scores distributions of classifiers (Gradient Boosting) trained on word embeddings derived from benchmarked Natural Language Processing (NLP) models. As per this assessment technique, the FTW2V model, trained from scratch with limited computational resources, outperforms generic contextual models in terms of word embeddings quality. Additionally, a computationally efficient contextual model rooted in FTW2V is introduced. This modified model substitutes Gradient Boosting with a transformer and integrates Part Of Speech labels.**

***Index Terms*—Natural Language Processing, Word embeddings, Gradient Boosting, hepatic, surgery, transformers, classifiers, supervised learning**

I. Introduction and related works

A substantial portion of the language models employed in the field of Natural Language Processing (NLP) serve to generate word embeddings. These embeddings can be tailored to various segments of the corpus, contingent upon the tokenization approach adopted by the NLP model. This may encompass sub-word vectorizations, individual words, or even entire sentences. The derived word embeddings typically serve as intermediary components within a given use case and can subsequently find application across diverse NLP endeavors, such as similarity analysis, topic identification, classification, and more.

For instance, the research efforts encompassed in Spacy's French models [1] and CamemBERT [2] revolve around advancing natural language processing capabilities for the French language. Both resources entail the fine-tuning of initial models to better cater to the intricacies of French text. However, it's important to acknowledge that this refinement process necessitates significant computational resources due to the complexities inherent to the language and the model. Nonetheless, it's noteworthy that the vocabulary upon which these models are trained tends to lean towards the general domain rather than being specialized, which can have implications for their performance in specific domains or industries.

Fine-tuning contextual NLP models often demands substantial data and computational resources [3]. When such resources are limited, alternatives to large model fine-tuning must be explored. This paper presents a resource-efficient method, FTW2V[1], which combines FastText and Word2Vec[2] embeddings. It's distinct from meta-embeddings. While meta-embeddings [4] combine various embeddings using techniques like concatenation, FTW2V uses FastText as an intermediary for Word2Vec embeddings, particularly for spelling errors or OOV words. This ensures optimal word embeddings for queried words.

Since this paper introduces novel embeddings and conducts a comparative analysis of their semantic relevance against pre-trained generic models, it also delves into the crucial question of how to assess the performance of an embedding and determine its optimal applicability within specific contexts (task scope and typology). The study explores diverse approaches to evaluating word embeddings for a given subdomain, as discussed by [5]. These approaches encompass both qualitative and quantitative methods, including:

- **Analogy**: This method involves analogical reasoning, where a tested word is substituted in a given analogy, and the objective is to predict the second word of the pair. For instance, if the original pair is (brother-sister), the task might be to predict the second word when the tested word becomes "grandson" (the expected answer being "granddaughter") [6].

[1]The model resulting from the combination of Word2Vec and FastText is denoted FTW2V

[2]In all this paper, when we write Word2Vec, we refer to its CBOW version

- **Similarity**: This approach relies on datasets comprising pairs of similar words or sentences. The cosine similarity between embeddings is calculated and then contrasted with human-annotated ground truth data. This method aims to measure the proximity of word meanings based on the embeddings.
- **Quality of Unsupervised Clustering**: In this method, unsupervised clustering is employed to identify underlying topics within a dataset. The qualitative assessment involves evaluating whether the identified clusters correspond coherently to relevant topics, providing insights into the embedding's ability to capture semantic relationships.

These evaluation techniques offer diverse perspectives for gauging the effectiveness of word embeddings within a subdomain, catering to both qualitative understanding and quantitative analysis.

There are also studies [7] that provide benchmarks between different language models (Word2Vec [8], FastText [9]...) and prove, for example, that Word2Vec and GloVe [10] yield better results than FastText. [11] also uses combination of static and contextual embeddings but for Named Entity Recognition (NER) instead of sentence classification, and in English only.

With regard to NLP models, especially generative models, model evaluation can also rely on the evaluation of perplexity and its evolution [12]. Other approaches are also introduced to evaluate the learning quality of a NLP model [13].

To summarize, the main questions that this paper studies are:

1) How can we identify word embeddings that are already promising candidates for fine-tuning within a specialized domain (hepatic surgery in the French language)?
2) Would a word embedding provided by the novel approach proposed in this paper (FTW2V), trained from scratch with limited computational resources and on a restricted corpus, exhibit enhanced performance in the context of the hepatic surgery subdomain?
3) Would the utilization of a classifier based on a transformer architecture, in lieu of Gradient Boosting, using the same FTW2V embedding, result in significant improvements in binary sentence classification scores?

The selection of the specific domain for testing these novel methods was influenced by Chaire BOPA, a French medical innovation center focused on hepatic surgery. Chaire BOPA initiated the study and supplied the input data for the research.

The paper primarily focuses on addressing the first and second questions with an opening to the third question. The paper is structured into the following key sections:

1) Introduction and related works
2) Approach and experiments: This section offers comprehensive insights into the overall approach, input data, preprocessing techniques, and the models used for benchmarking
3) Results and discussions: This part presents the results related to the effectiveness of word embeddings and the advantages associated with the utilization of a transformer classifier
4) Conclusion and future work

II. Approach and Experiments

A. General description of the approach

Word embedding evaluation methods [14] fall into two categories: intrinsic and extrinsic. Intrinsic methods directly assess embeddings through linguistic or similarity tasks. Extrinsic methods gauge performance by utilizing word embeddings as inputs for task-specific models.

In this paper, the proposed approach for evaluating the relevance of word embeddings, whether from pre-trained or retrained models, within a specialized corpus (hepatic surgery) relies on classification scores derived from elementary classification tasks. Notably, the area under the precision-recall curve (AUC-PR) is employed. A pre-trained model's word embedding is deemed suitable for a specialized domain if the sentence's embedding enables the classifier to generate intermediate features that facilitate effective discrimination, leading to favorable performance in elementary classification tasks. Essentially, if the provided word embedding lacks semantic relevance, classifiers trained with it cannot construct intermediate decision trees conducive to successful classification. Assessing multiple classification tasks, as opposed to a single one, enhances the robustness of conclusions by addressing various semantic facets inherent to the considered domain.

The key steps of the developed approach are as follows:

- Collecting a corpus of French medical reports pertinent to hepatic surgery;
- Identifying classification tasks that will serve as benchmarks for classifiers;
- Pre-processing and cleansing the corpus;
- Retrieving word-level word embeddings provided by the NLP model under assessment and computing the mean embeddings per sentence;
- Training a Gradient Boosting model for each classification task, employing the sentence-level word embeddings;
- Assessing classifiers on a test dataset concerning each of the classification tasks;
- Aggregating outcomes and computing performance metrics.

For each sentence classification task, precision, recall, and AUC-PR are assessed on a test dataset through multiple random samplings. Subsequently, a mean score is computed for each model.

B. Input Data

In this work, a classifier mainly consists of two components: the Natural Language Processing (NLP) model that generates word embeddings, which is shared across different tasks, and the classification algorithm (here Gradient Boosting) that is trained for each task but takes as input word embeddings provided by the NLP models. These two components require two different types of training data:

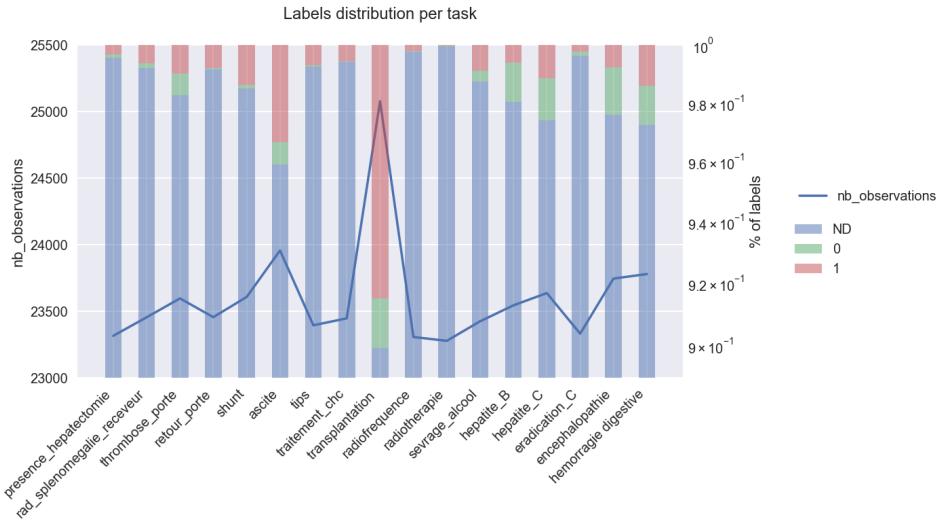

Fig. 1. Number of annotated sentences per classification task and distribution of labels.

Classification tasks	ND	1	0	Number of observations
presence_hepatectomie (*presence_hepatectomy*)	0.9955	0.0032	0.0012	23314
rad_splenomegalie_receveur (*rad_splenomegaly_receiver*)	0.9919	0.0065	0.0016	23453
thrombose_porte (*thrombosis_gate*)	0.9825	0.0099	0.0076	23595
retour_porte (*back_door*)	0.9914	0.0081	0.0005	23454
shunt	0.9850	0.0139	0.0011	23606
ascite (*ascites*)	0.9590	0.0336	0.0074	23955
tips	0.9923	0.0071	0.0006	23393
traitement_chc (*chc treatment*)	0.9942	0.0056	0.0003	23445
transplantation	0.8992	0.0850	0.0158	25077
radiofrequence (*radio frequency*)	0.9976	0.0021	0.0003	23305
radiotherapie (*radiotherapy*)	0.9995	0.0003	0.0002	23276
sevrage_alcool (*alcohol withdrawal*)	0.9872	0.0090	0.0038	23423
hepatite_B (*Hepatitis B*)	0.9802	0.0062	0.0135	23543
hepatite_C (*Hepatitis C*)	0.9740	0.0116	0.0143	23635
eradication_C	0.9963	0.0024	0.0013	23330
encephalopathie (*encephalopathy*)	0.9758	0.0079	0.0163	23744
hemorragie_digestive (*digestive bleeding*)	0.9723	0.0141	0.0136	23778

TABLE I
NUMBER OF ANNOTATED SENTENCES PER CLASSIFICATION TASK AND DISTRIBUTION OF LABELS.

- Unlabeled data consisting of specialized corpora extracted from various sources by surgeons. These data are solely used to train the NLP models from scratch using the new combination algorithms (FTW2V) proposed in this paper (see Fig. 2).
- Annotated and anonymized data consisting of sentences or paragraphs annotated with respect to binary questions (classification tasks) to train the Gradient Boosting models with NLP models word embeddings as input features. Annotated data is also used to assess the classification performance on a test set that is different from the one used for training.

APHP (Greater Paris University Hospitals), collaborating with Chaire BOPA specialized in hepatic surgery innovation, provides both labeled and unlabeled data. Medical experts have annotated the labeled data. The data utilized in this study

are not available to the general public due to privacy and intellectual property concerns.

The characteristics of unlabeled data are as follows:

- Number of sources: 5 (either database extractions or retrieved medical reports, but all relating to the same medical specialty, namely hepatic surgery).
- Number of paragraphs in the corpus: 157,772.
- Number of words: 32,374,012.

The sentence annotation for the second dataset was performed by medical experts who assigned one of the following labels to the annotated sentences extracted from medical reports:

- ND: Not defined, which means that the sentence is off-topic with respect to the question asked.
- 0: No or absent.
- 1: Yes or present.

During the classifier training phase, and in order to reduce it to a simple binary classification case, the two labels ND and 0 were considered as belonging to the same label (0). The new resulting classification task consists more of telling whether the sentence confirms the presence of the pathology.

A total of 17 classification tasks were addressed, each linked to a distinct binary question detailed in Table I. Specifically, each classification task involves assigning a response of 1 (indicating presence) or 0 (indicating absence) to its corresponding question. The list of the 17 questions is provided in Table I and also depicted in Fig. 1, along with the distribution of annotated sentence labels.

For instance, referring to Table I, in the context of the classification task linked to the *presence_hepatectomy* label, a total of 23,314 sentences were annotated. Among these, 99.55% were labeled as ND, 0.32% as 0, and 0.12% as 1.

All data sets for the classification tasks exhibit imbalanced distributions. However, certain tasks display a relatively higher proportion of observations with label 1 compared to others, potentially accounting for the variance in classifier performance across different tasks.

C. Pre-processing

The same preprocessing was applied to both annotated and unlabeled corpora, aiming to condense vocabulary size. The steps encompass:

- Removal of special characters;
- Conversion of sentences to lowercase;
- Retention of punctuation, contributing to task-specific meaning;
- Lemmatization utilizing the spaCy fr_core_news_md model [1].

D. Compared NLP models

Among the benchmarked models, spaCy models are present. spaCy is a software library that delivers diverse language models and modular NLP components for various languages [15]. While the architecture of these models is undisclosed, their website provides relevant information [1].

MedSpaCy [16], an adaptation of spaCy for medical contexts, is also included. However, designed for English, MedSpaCy necessitated translation using the Python package "deep-translator."

CamemBERT [17] is a French language model rooted in RoBERTa. Among its 6 available versions on HuggingFace, this paper focuses on the main and generic variants.

For evaluating diverse word embedding strategies, we also incorporate CharBERT [18]. CharBERT differs from BERT by utilizing character-level embeddings, circumventing BERT's WordPiece tokenization limitations [18].

The generic pre-trained models included in the benchmark and used to provide word embeddings to classifiers are as follows:

- CamemBERT-base [2] (denoted CB_base).
- CamemBERT-large [2] (denoted CB_large).
- spaCy fr_core_news_md [1] (denoted SP_md).
- spaCy fr_core_news_lg [1] (denoted SP_lg).
- MedSpaCy en_core_sci_lg (denoted SCISP_lg).
- CharBERT medical_character_bert (denoted CHARBERT).

These models are compared against the novel approach introduced in this paper: NLP models trained from scratch on a specialized unlabeled corpus, generating word embeddings. This new approach comes in two versions:

- FTW2V: Combination of Word2Vec [8] and FastText [9].
- FTGloVe: Combination of GloVe [10] and FastText [9] [3].

To verify that FTW2V performs at least on par with FastText or W2V alone, the embeddings of these two models, trained from scratch on the same corpora, are incorporated into the benchmark:

- Word2Vec alone [8] (denoted W2V).
- FastText alone [9] (denoted FT).

While the combination of FastText and Word2Vec in the FTW2V model may not substantially elevate classification scores, particularly if the test datasets lack spelling errors, it is employed to address situations where Word2Vec's vocabulary lacks certain words. This capability is achieved through the algorithm outlined in Fig. 2.

This approach's (Fig. 2) efficacy is augmented by furnishing consistent word embeddings. The algorithm consistently derives word embeddings from Word2Vec (or GloVe), bypassing FastText for this purpose. FastText's role lies in establishing word relationships based on their constituent characters. The algorithm then utilizes Word2Vec's (or GloVe's) embedding of the closest word, as per FastText, for returning the word embedding. This algorithm yields two key advantages:

- Provision of word embeddings for terms absent in Word2Vec's or GloVe's vocabulary.

[3]The primary focus in this paper is on FTW2V due to its comparable performance with FTGloVe and the potential errors in the GloVe python package.

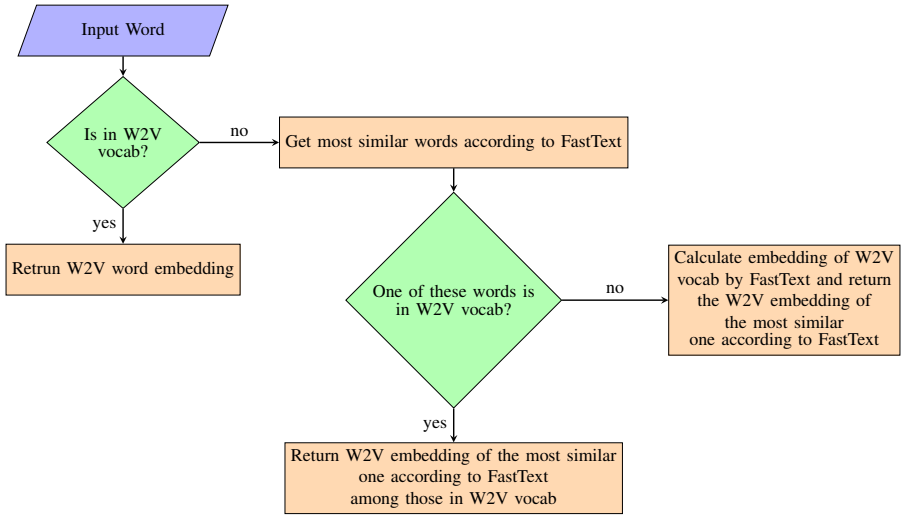

Fig. 2. Combination of Word2Vec (GloVe) with FastText

- Effective management of spelling errors.

E. Training and Evaluation

1) Comparison between word embeddings: Given imbalanced data, Gradient Boosting [19] was chosen over Random Forest as the learning model. NLP model embeddings served as input features. Boosting iteratively emphasizes mispredicted instances, while RandomForest forms an ensemble through Bootstrap Aggregating. Boosting's focus on misclassification typically outperforms RandomForest's random sampling for imbalanced data, supported by literature [20].

Stratified datasets maintained label distribution (0 or 1) between train and test datasets. Each classifier underwent training, estimating Precision, Recall, and AUC-PR. AUC-PR is better suited for imbalanced data than ROC AUC [21] [22]. Other valid metrics like F-scores or weighted area under the Recall-Precision curve [23] aren't considered due to their lower prevalence.

Evaluation on the different classification tasks allows each classifier to have 17 estimated metric values, and the distribution of the results obtained is provided in the following sections.

2) Use of a transformer instead of Gradient Boosting: With the most relevant word embedding established in the previous subsection, the potential for performance enhancement emerges by replacing Gradient Boosting with an alternative classifier.

This approach introduces a new classifier (ATFTW2V) , based on transformers. Sentences are represented using a 50 x 355 matrix, comprising word vectors from FTW2V, spaCy's Part-Of-Speech (POS) tags, and other encoded labels. Adding POS tags aligns with findings [24] that underscore their performance impact.

The same training strategy is applied with a trained classifier to each classification task. Comparison with the Gradient Boosting classifier in the next section reveals an improvement in performance.

III. Results and Discussions

A. Comparison of word embeddings relevance

Pre-trained NLP models, along with classifiers trained for identical classification tasks on the same training data, yielded precision, recall, and AUC-PR scores outlined in Table II. Notably, the distribution of the most significant score, AUC-PR, is presented in Fig. 3.

The primary metric in this study is the AUC-PR score, offering threshold-independent performance assessment. Gradient Boosting classifiers with NLP model embeddings FTW2V and FTGLOVE outperform those using "generic" pre-trained models.

Notably, spaCy's models match AUC-PR scores of CamemBERT (base and large) (CB_large and CB_base), with reduced variation. Additionally, SCISP_lg (MedSpaCy) outperforms despite a generic translation step, emphasizing specialized model potential.

Results also highlight FTW2V's edge over FT and W2V, tempered by limited spelling errors in test data. This why in III-B2 specific examples showing the strength of combining FastText and W2V are provided.

B. Improving performance by using a different classifier

1) New transformer-based approach compared to Gradient Boosting: In the previous section, it was shown that word embeddings of FTGLOVE or FTW2V trained from scratch on a specialized corpus give more relevant word embedding than the generic models considered in this benchmark.

In this section, the objective is to evaluate whether it is possible to improve performance by replacing Gradient Boosting with the approach described in II-E2 using the same word embeddings. Results obtained are detailed in Fig. 4 and in Table III.

The performance distinction is notable, especially concerning recall. This indicates that transformers effectively leverage location-based similarities compared to an approach relying

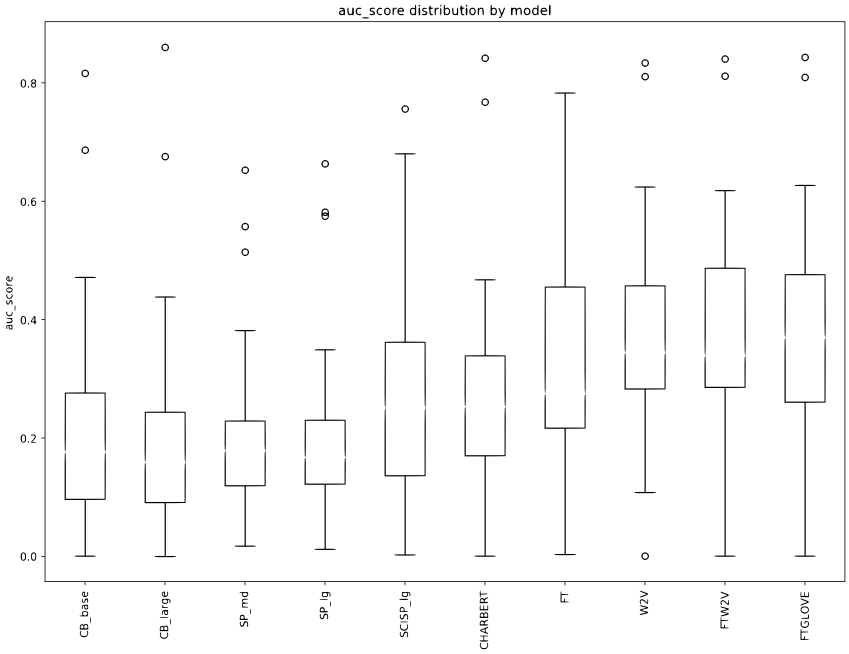

Fig. 3. AUC-PR distribution - Pretrained generic models vs trained from scratch

	CB_base	CB_large	SP_md	SP_lg	SCISP_lg	CHARBERT	FT	W2V	FTW2V	FTGLOVE
mean_auc-pr	0.24	0.22	0.23	0.23	0.30	0.29	0.34	0.38	**0.39**	0.38
std_auc-pr	0.22	0.23	0.19	0.20	0.22	0.23	0.23	0.23	0.23	0.23
mean_recall	0.25	0.23	0.16	0.15	0.21	0.28	0.26	**0.31**	**0.31**	**0.31**
std_recall	0.15	0.15	0.13	0.11	0.13	0.16	0.19	0.16	0.17	0.17
mean_precision	0.38	0.34	0.40	0.39	0.48	0.42	0.49	0.53	**0.54**	0.53
std_precision	0.24	0.25	0.28	0.27	0.27	0.24	0.26	0.22	0.24	0.24

TABLE II
DISTRIBUTION OF MODEL SCORES - PRETRAINED MODELS VS TRAINED FROM SCRATCH

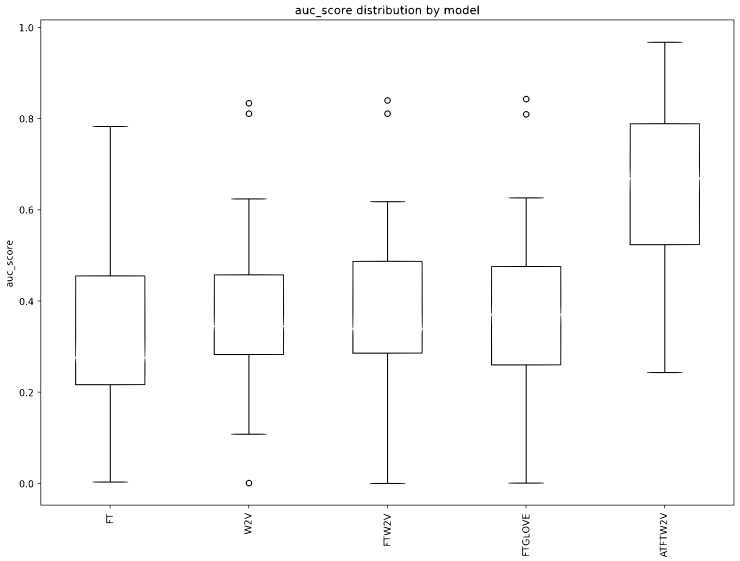

Fig. 4. AUC-PR distribution - Transformers vs Gradient Boosting

	FT	W2V	FTW2V	FTGLOVE	ATFTW2V
mean_auc-pr	0.34	0.38	0.39	0.38	**0.67**
std_auc-pr	0.23	0.23	0.23	0.23	0.20
mean_recall	0.26	0.31	0.31	0.31	**0.68**
std_recall	0.19	0.16	0.17	0.17	0.24
mean_precision	0.49	0.53	0.54	0.53	**0.64**
std_precision	0.26	0.22	0.24	0.24	0.22

TABLE III
DISTRIBUTION OF MODELS SCORES - TRANSFORMERS (ATFTW2V) VS GRADIENT BOOSTING

solely on a decision tree ensemble like Gradient Boosting. Additional performance improvements can be attained through hyperparameter optimization.

2) Analysis of some inference examples: The following examples provide insights into the performance and behavior of a specific classifier introduced in section II-E2. In each example, two versions of a sentence are presented with minor variations to evaluate the classifier's evolution and robustness in predicting scores. The predictions are obtained by submitting these sentences to the model and analyzing the resulting scores for different classification tasks. By examining scenarios involving spelling errors, synonyms, and negation, we can gain a deeper understanding of the classifier's capabilities and limitations. The predictions obtained with different word embeddings, namely FTW2V and W2V, are compared to assess the impact of these embeddings on the model's performance. Through these examples, we can uncover important insights into the behavior of the classifier and its sensitivity to various linguistic factors.

Example 1: Robustness to Spelling Errors The following example demonstrates the classifier's robustness in handling spelling errors, using the classifier introduced in section II-E2.

- Version 1 of the sentence in French: Le patient représente un **saignement** important (*The patient represents a major bleeding*).
- Model prediction with FTW2V: The classification task "digestive bleeding" received a score of 0.9, while the other classification tasks received a score of 0.
- Model prediction with W2V: The classification task "digestive bleeding" received a score of 0.9, while the other classification tasks received a score of 0.
- Version 2 of the sentence in French: Le patient représente un **saiggnement** important (a spelling error on "saiggnement," with 2 'g's instead of one, *The patient represents a major bleedding*).
- Model prediction with FTW2V: The classification task "digestive bleeding" received a score of 0.9, while the other classification tasks received a score of 0. This prediction seems consistent.
- Model prediction with W2V: All classification tasks, including "digestive bleeding," received a score of 0.0.
- Conclusion: The spelling error in the term "saignement" (*bleeding*) does not appear to have any impact on the results with FTW2V. This robustness is achieved through the combination of FastText and Word2Vec embeddings. However, when W2V alone was assessed, the prediction for all classes was 0. In this case, the main meaning of the word was not recognized due to the spelling error.

Example 2: Robustness to Synonyms The following example demonstrates the classifier's ability to handle synonyms and highlights the relevance of word embeddings in this context.

- Version 1 of the sentence in French: Le patient a eu une **transplantation** (*The patient underwent transplantation*).
- Model prediction with FTW2V: The scores were 0.01 for Hepatitis B, 0.78 for transplantation, and 0 for all other classes. This prediction seems consistent.
- Version 2 of the sentence in French: Le patient a eu une **greffe de foie** (*The patient had a liver transplant*).
- Model prediction with FTW2V: The score for transplantation was 0.75, and 0 for all other questions. This prediction seems consistent.
- Conclusion: This example illustrates the model's ability to associate different but semantically similar terms.

Example 3: Evolution of Scores with Negation The following example illustrates the evolution of scores when negation is introduced, using the classifier introduced in section II-E2.

- Version 1 of the sentence in French: Le patient **a eu** une greffe de foie (*The patient had a liver transplant*).
- Model prediction with FTW2V: The score for transplantation was 0.75, and 0 for all other questions. This prediction seems coherent.
- Version 2 of the sentence in French: Le patient **n'a pas eu** de greffe de foie (*The patient did not have a liver transplant*).
- Model prediction with FTW2V: The score for transplantation was 0.46, and 0 for all other questions. This prediction seems consistent.
- Conclusion: Introducing negation in the sentence resulted in a lower score, which aligns with logical expectations. However, the obtained score of 0.46, although below the decision threshold of 0.5, remains relatively high. This suggests that the model's performance in handling negation could be further improved. Nonetheless, the prediction remains consistent in terms of indicating a lower likelihood of transplantation in the negated sen-

tence compared to the affirmative sentence.

In this example, we observe the classifier's response to the introduction of negation and how it affects the prediction scores. The decrease in the score for the transplantation task indicates the model's understanding of the negated statement. However, further refinements may be necessary to achieve more satisfactory results and reduce the relatively high score obtained in the negated sentence.

IV. Conclusion and Future Work

In this study, our focus has been on devising inventive strategies to create robust and resource-efficient model that effectively handle out-of-vocabulary words and provide word embeddings that are relevant for a specialized domain. A novel framework is introduced within this paper, which merges the strengths of Word2Vec or GloVe with FastText, resulting in the development of FTW2V embeddings.

Throughout our research, we've underscored the paramount importance of domain-specific training in producing word embeddings that carry semantic relevance. The limitations of generic models have been highlighted, showcasing the need for more tailored approaches. Our study has demonstrated that efficient methods can yield highly promising results, even outperforming pre-trained models in benchmark tests. By integrating transformers, we've effectively fused static models with FastText's capacity to handle errors, thus enhancing classification tasks.

The paper also presents a novel method for assessing relevance, based on estimating the distribution of AUC-PR scores on several classification tasks. This approach contributes to a more comprehensive evaluation of embedding quality.

Moving forward, applying a similar approach to different specialized fields would necessitate possessing unlabeled data on a comparable scale to what was used in this study. By utilizing this data, we could train Word2Vec and FastText models. These models, when integrated into the algorithm depicted in Fig. 2, would lead to the development of the FTW2V model. Subsequently, the embeddings generated by FTW2V could be employed across various natural language processing tasks relevant to the specific domain.

Future endeavors include delving into different transformer architectures, optimizing hyper-parameters, and crafting computationally-efficient models for customized Named Entity Recognition tasks. We also foresee adapting this method to diverse domains.

References

[1] spacy, "SpaCy french models," https://spacy.io/models/fr, 2022.

[2] F. A. Research, Inria, and ALMAnaCH, "camemBERT," https://camembert-model.fr/, 2022.

[3] D. Vucetic, M. Tayaranian, M. Ziaeefard, J. J. Clark, B. H. Meyer, and W. J. Gross, "Efficient fine-tuning of bert models on the edge," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022, pp. 1838–1842.

[4] D. Bollegala and J. O'Neill, "A survey on word meta-embedding learning," in *IJCAI: proceedings of the conference/sponsored by the International Joint Conferences on Artificial Intelligence*, 2022.

[5] Y. Yaghoobzadeh, K. Kann, and H. Schütze, "Evaluating word embeddings in multi-label classification using fine-grained name typing," in *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018, pp. 101–106.

[6] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of ACL (Volume 1: Long Papers)*, 2014, pp. 238–247.

[7] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA transactions on signal and information processing*, vol. 8, p. e19, 2019.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of EMNLP*, 2014, pp. 1532–1543.

[11] H. El Boukkouri, O. Ferret, T. Lavergne, and P. Zweigenbaum, "Embedding Strategies for Specialized Domains: Application to Clinical Entity Recognition," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019.

[12] H. Ngo, J. G. Araújo, J. Hui, and N. Frosst, "No news is good news: A critique of the one billion word benchmark," in *arXiv preprint arXiv:2110.12609*, 2021.

[13] C. Meister and R. Cotterell, "Language model evaluation beyond perplexity," in *Proceedings of ACL (Volume 1: Long Papers)*, 2021, pp. 5328–5339.

[14] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA transactions on signal and information processing*, vol. 8, p. e19, 2019.

[15] S. Tual, N. Abadie, J. Chazalon, B. Duménieu, and E. Carlinet, "A benchmark of nested named entity recognition approaches in historical structured documents," *arXiv preprint arXiv:2302.10204*, 2023.

[16] H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson, "Launching into clinical space with medspacy: a new clinical text processing toolkit in python," in *AMIA Annual Symposium Proceedings*, vol. 2021. American Medical Informatics Association, 2021, p. 438.

[17] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[18] W. Ma, Y. Cui, C. Si, T. Liu, S. Wang, and G. Hu, "Charbert: Character-aware pre-trained language model," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 39–50.

[19] S. Fafalios, P. Charonyktakis, and I. Tsamardino, "Gradient boosting trees," 2020.

[20] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012.

[21] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM computing surveys (CSUR)*, vol. 49, no. 2, pp. 1–50, 2016.

[22] J. Frery, "Ensemble learning for extremely imbalanced data flows," Ph.D. dissertation, Université de Lyon, 2019.

[23] C. K. Williams, "The effect of class imbalance on precision-recall curves," *Neural Computation*, vol. 33, no. 4, pp. 853–857, 2021.

[24] A. Benamar, M. Bothua, C. Grouin, and A. Vilnat, "Easy-to-use combination of pos and bert model for domain-specific and misspelled terms," in *NL4IA Workshop Proceedings*, 2021.