Neural Text Classification and Stacked Heterogeneous Embeddings for Named Entity Recognition in SMM4H 2021

Usama Yaseen[1,2], Stefan Langer[1,2]
[1]Technology, Siemens AG Munich, Germany
[2]CIS, University of Munich (LMU) Munich, Germany
{usama.yaseen,langer.stefan}@siemens.com

Abstract

This paper presents our findings from participating in the SMM4H Shared Task 2021. We addressed Named Entity Recognition (NER) and Text Classification. To address NER we explored BiLSTM-CRF with Stacked Heterogeneous Embeddings and linguistic features. We investigated various machine learning algorithms (logistic regression, Support Vector Machine (SVM) and Neural Networks) to address text classification. Our proposed approaches can be generalized to different languages and we have shown its effectiveness for English and Spanish. Our text classification submissions (team:MIC-NLP) have achieved competitive performance with F1-score of 0.46 and 0.90 on ADE Classification (Task 1a) and Profession Classification (Task 7a) respectively. In the case of NER, our submissions scored F1-score of 0.50 and 0.82 on ADE Span Detection (Task 1b) and Profession Span detection (Task 7b) respectively.

1 Introduction

The ubiquity of social media has led to massive user-generated content across various platforms. Twitter is a popular micro-blogging platform that allows its users to publish tweets up to 280 characters. The common public uses Twitter to share life-related personal and professional experiences with others. Personal experiences often involve health-related incidents including mentions of adverse drug effect (ADE); this information is crucial to study Pharmacovigilance. In the context of the COVID-19 pandemic, the professional experiences may include information about professions and occupations which are vulnerable due to either direct exposure to the virus or due to the associated mental health issues; detecting vulnerable occupations is critical to adopt necessary preventive measures.

Recent research focuses on mining Twitter data for adverse drug effect detection (Jiang and Zheng, 2013; Adrover et al., 2015; Onishi et al., 2018). The distinctive style of communication on Twitter presents unique challenges including informal (brief) text, misspellings, noisy text, abbreviations, data sparsity, colloquial expressions and multilinguality.

2 Task Description and Contribution

We participate in the following two tasks organized by SMM4H workshop 2021 (Magge et al., 2021): (1) Task 1: Classification, Extraction and Normalization of Adverse Effect mentions in English tweets (2) Task 7: Identification of professions and occupations in Spanish tweets (Miranda-Escalada et al., 2021). Task 1 consists of three sub-tasks, (a): ADE tweet classification, (b): ADE span detection, (c): ADE resolution; whereas Task 7 consists of two sub-tasks: (a): Tweet classification (b): Profession/occupation span detection. For both tasks, we participate in sub-tasks (a) and (b). The Task 1a and Task 7a is a text classification problem while Task 1b and Task 7b is a Named Entity Recognition problem.

Following are our multi-fold contributions:

1. To address NER tasks, we have employed a neural network based sequence classifier, i.e. BiLSTM-CRF and investigated various heterogeneous embeddings. We further investigated the combination of character embeddings, static word embeddings and contextualized embeddings in a stacked format. We also incorporated linguistic features such as part-of-speech tags (POS), orthographic features etc. We apply the proposed modelling approaches to both English and Spanish texts. In *Profession span detection* (Task 7b) our submission (team:MIC-NLP) achieved the F1-score of 0.824 which is 6 points higher than the arithmetic median of all the submissions; in case of *ADE span detection* our submission scored F1-score of 0.50, around 8 points higher than the arithmetic median of the participating submissions.

2. To address text classification tasks, we investi-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 83–87

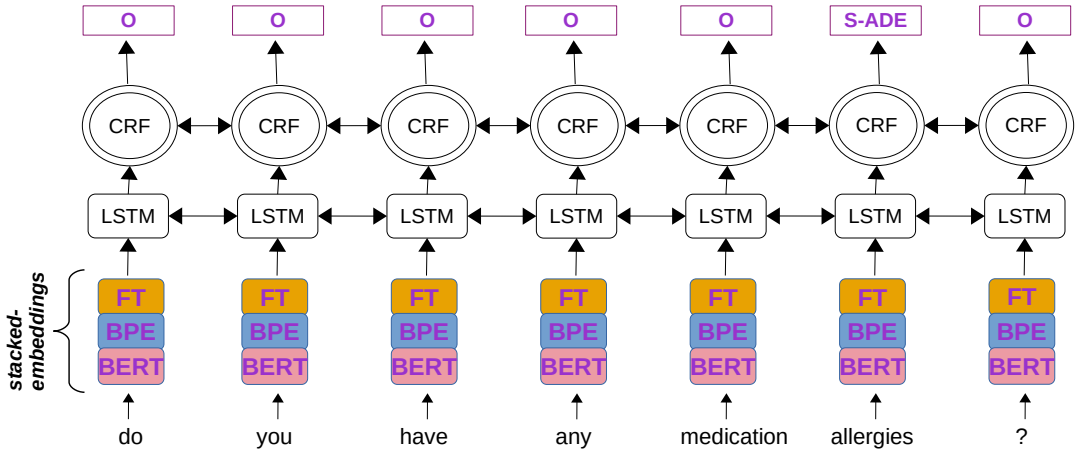

Figure 1: System architecture for NER task, consisting of BiLSTM-CRF with stacked heterogeneous embeddings. Here, *FT*: fastText embedding vector; *BPE*: Byte-Pair embedding vector; *BERT*: BERT embedding vector; *S_ADE*: S_Adverse Drug Effect.

gated various machine learning algorithms like *logistic regression*, *SVM* and *neural network* with various word and sentence embeddings. In *ADE tweet classification* (Task 1a) our submission (team:MIC-NLP) scored F1-score of 0.46, approximately 2 points higher than the arithmetic median of participating submissions; in case of *tweet classification* (task 7a) our system achieved the F1-score of 0.90 which is 5 points higher than the arithmetic median of all submissions.

3 Methodology

In the following sections we discuss our proposed model for named entity recognition and text classification.

3.1 Named Entity Recognition

Figure 1 describes the architecture of our model, where we design a sequence tagger to extract entities. The architecture of our model is a standard BiLSTM-CRF (Lample et al., 2016) model with stacked heterogeneous embeddings and linguistic features as input. The stacked embeddings consists of Byte-Pair subword embeddings (Heinzerling and Strube, 2018), fastText subword embeddings (Bojanowski et al., 2017) and contextualized word embeddings (Devlin et al., 2019; Liu et al., 2019). The linguistic features include POS, capitalization features and orthographic features.

3.2 Text Classification

We explored traditional machine learning algorithms like logistic regression, SVM and neural network based architecture with various word and sentence embeddings for text classification. The SVM was trained with Radial Basis Function (RBF) Kernel with the value of penalty parameter C determined by grid search for each dataset. Our best model was a Neural Network with contextualized embeddings (Devlin et al., 2019; Liu et al., 2019). Since both datasets (Task 1a and Task 7a) were highly imbalanced, we employed higher class weights for minority classes to train the final models.

Task	Train	Dev
Sentence Counts		
Task 1b	34142	1775
Task 7b	14755	4959
Task 1b Entities		
ADE	1713	87
Task 7b Entities		
PROFESION	1597	566
SITUACION_LABORAL	264	85
ACTIVIDAD	45	16
FIGURATIVA	16	8

Table 1: Dataset statistics for NER.

3.3 Ensemble Strategy

Bagging is a useful technique to reduce the variance of the learning algorithm without impacting bias. We employed a variant of Bagging (Breiman, 1996) such that every data point in the training set is part of the development set at least once and vice versa. We created three data folds and trained the model using optimal configuration on each fold, inference on the test set involves majority voting among the

Hyper-parameter	Value
NER	
learning rate	0.1
optimizer	SGD
hidden size	256
POS dimensions	50
Ortho dimension	50
batch size	32
epochs	150
Text Classification	
kernel	RBF
class-weights	10.0
learning rate	0.00003
batch size	16
epochs	10

Table 2: Hyper parameter settings for NER and Text classification.

three trained models.

For NER, we perform majority voting at the token level for each test data point. In cases when voting results in a tie, we take the prediction of the confident model, we treat the model trained on original data split as the confident model. In the case of an ensemble for text classification, we followed the straight forward approach of majority voting at sentence level for each test data point.

4 Experiments and Results

4.1 Dataset and Experimental Setup

Data: We employed bagging (discussed in section 3.3) to split the annotated corpus into 3-folds. For ADE span detection (Task 1b) and Profession span detection (Task 7b) we perform sentence splitting, word tokenization, computing orthographic features and POS tagging. We do not perform any pre-processing for ADE classification (Task 1a) and Tweet classification (Task 7a).

ADE Classification (Task 1a): The dataset consists of tweets in the English language and the task is to detect tweets containing adverse drug effect. The dataset contains two classes, *ADE* and *NoADE*. The dataset is highly imbalanced with only 1235 tweets of type ADE out of total 17385 tweets in the train set.

ADE Span Detection (Task 1b): The dataset consists of only one entity type *ADE*. The train set contains 1717 entity mentions of *ADE* (see Table 1).

	Features	Task 1b P/R/F1	Task 7b P/R/F1
r1	*glove*	.5/.18/.26	-
r2	*fastText*	.89/.28/.43	.84/.64/.73
r3	*fastText + Char*	.64/.28/.39	.83/.67/.74
r4	*fastText + BytePair*	.62/.34/.44	.82/.74/.78
r5	*BERT*	.68/.35/.46	.84/.76/.80
r6	*BERT + fastText + BytePair*	**.61/.52/.56**	**.86/.77/.81**
		Fold=2	**Fold=2**
r7	*BERT + fastText + BytePair*	**.80/.21/.34**	**.85/.79/.82**
		Fold=3	**Fold=3**
r8	*BERT + fastText + BytePair*	**.77/.37/.50**	**.84/.78/.81**

Table 3: Scores on dev set using different features for *BiLSTM-CRF* on *Task 1b* and *Task 7b*.

Profession Classification (Task 7a): The dataset consists of tweets in the Spanish language and the task is to detect tweets containing mention of profession/occupation. The dataset contains two classes. The dataset is highly imbalanced with only 1393 tweets containing a positive mention out of 6000 tweets.

Profession Span Detection (Task 7b): The dataset consists of four entity types with few mentions of type *FIGURATIVA* as shown in Table 1. Entities of type ACTIVIDAD and FIGURATIVA are ignored in the evaluation of this shared task but we still treat them as regular entities.

Experimental Setup: We found contextualized embeddings to be very helpful in identifying entities and text classification; all our experiments used pre-trained contextualized embeddings. We employ *RoBERTa* (Gururangan et al., 2020) for Task 1a and Task 1b; we use multi-lingual BERT (Devlin et al., 2019) for Task 7a and Spanish BERT (Cañete et al., 2020) for Task 7b. We do not finetune embeddings in our experiments. We don't employ any strategy for handling imbalanced classes for NER but have used class weighting by a factor of 10 for all positive classes for text classification. Table 2 lists the best configuration of hyperparameters for all the tasks.

4.2 Results on Development Set

We perform various experiments to investigate the impact of features on performance on the development set.

NER: Table 3 shows the score on the development set for Task 1b and Task 7b. Observe that fastText embeddings (row r2) outperform glove embeddings (row r1) for Task 1b. Subsequently, fastText embeddings with BytePair embeddings (row r4) provide an improvement over only fast-

	Features	Task 1a P/R/F1	Task 7a P/R/F1
r1	*logisticReg + fastTextSentEmb*	.33/.83/.47	.38/.95/.55
r2	*logisticReg + BERTSentEmb*	.34/.81/.48	.41/.83/.55
r3	*logisticReg + BERTWordEmbSum*	.45/.86/.59	.45/.86/.59
r4	*SVM + fastTextSentEmb*	.53/.66/.59	.71/.67/.69
r5	*SVM + BERTSentEmb*	.36/.86/.51	.49/.66/.56
r6	*SVM + BERTWordEmbSum*	.44/.90/.59	.61/.64/.63
r7	*NeuralNetwork + Glove*	.51/.63/.56	.64/.59/.61
r8	*NeuralNetwork + BERT*	**.77/.72/.74**	**.95/.85/.90**
r9		**Fold=2**	**Fold=2**
r10	*NeuralNetwork + BERT*	**.79/.66/.72**	**.89/.91/.90**
r11		**Fold=3**	**Fold=3**
r12	*NeuralNetwork + BERT*	**0.8/.65/.72**	**.93/.84/.88**

Table 4: Scores on dev set using different features on *Task 1a* and *Task 7a*.

Text (row r2) and the combination of fastText with Character embeddings (row r3). The contextualized embeddings (row r5) provide an improvement over the combination of fastText with BytePair embeddings. In row r6, we employ BERT, fastText and BytePair embeddings in a stacked format leading to the best f1-score for both Task 1b and Task 7b.

Text Classification: Table 4 shows the score on the development set for Task 1a and Task 7a. Observe that BERTSentEmb provides improvement over fastTextSentEmb for both logistic regression and SVM. Similarly, BERTWordEmbSum further improves BERTSentEmb. BERTSentEmb uses BERT's *CLS* representation whereas BERTWordEmbSum is computed by average of the token-wise embeddings of pre-trained BERT as discussed in Rogers et al.. Neural Network with BERT achieves the best result for both datasets.

4.3 Results on Test Set

Table 5 shows the comparison of our submissions with the arithmetic median of the participating teams for all the tasks. Our submissions achieve the overall best F1-score than the arithmetic median for all the tasks showing compelling advantage. For Task 1a, the precision of our system is lower than the arithmetic median but this is compensated by the improvement in recall. For all the tasks, the precision is higher than the recall but overall precision and recall are balanced.

5 Conclusion

In this paper, we described our system with which we participate in Task 1(Adverse Drug Effect Classification and Extraction) and Task 7 (Identification of professions and occupations in Spanish Tweets) in the SMM4H Shared Task 2021. Our NER system employed stacked heterogeneous embeddings to extract entities in English and Spanish text. Our NER system demonstrates a competitive performance with F1-score of 0.50 and 0.82 on ADE Span Detection (Task 1b) and Profession/Occupation span detection (Task 7b) respectively. Our text classification system employed contextualized embeddings with Neural Network as a classifier to achieve a competitive performance with F1-score of 0.46 and 0.90 on ADE Classification (Task 1a) and Profession/Occupation classification (Task 7a) respectively. In future, we would like to improve error analysis to further enhance our NER and text classification models.

	Tasks	Arithmetic Median *P/R/F1*	MIC-NLP *P/R/F1*
r1	*Task 1a*	**.50**/.40/.44	.47/**.45**/**.46**
r2	*Task 1b*	.49/.45/.42	**.55**/**.45**/**.50**
r3	*Task 7a*	.91/.85/.85	**.94**/**.85**/**.90**
r4	*Task 7b*	.84/.72/.76	**.85**/**.79**/**.82**

Table 5: Comparison of our system (team:MIC-NLP) with the arithmetic median of the participating teams. Scores on test set for Task 1a, Task 1b, Task 7a and Task 7b.

Acknowledgment

This research was supported by Bundesministerium für Wirtschaft und Energie (bmwi.de), grant 01MD19003E (PLASS, plass.io) at Technology - Siemens AG, Munich Germany.

References

Cosme Adrover, Todd J. Bodnar, Z. Huang, A. Telenti, and M. Salathé. 2015. Identifying adverse effects of hiv drug treatment and associated sentiments using twitter. volume 1.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. volume 5, pages 135–146.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Keyuan Jiang and Yujing Zheng. 2013. Mining twitter data for potential drug effects. In *ADMA*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. volume abs/1907.11692.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima López, Vicent Briva-Iglesias, Marvin Agüero-Torales, Luis Gascó-Sánchez, and Martin Krallinger. 2021. The profner shared task on automatic recognition of professions and occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Takeshi Onishi, Davy Weissenbacher, Ari Klein, Karen O'Connor, and Graciela Gonzalez-Hernandez. 2018. Dealing with medication non-adherence expressions in Twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 32–33, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Trans. Assoc. Comput. Linguistics*, 8:842–866.