hhuEDOS at SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS) – Binary Sexism Detection (Subtask A)

Wiebke Petersen and **Diem-Ly Tran** and **Marion Wroblewitz**
Heinrich Heine University Düsseldorf, Germany
wiebke.petersen@hhu.de, diem-ly.tran@hhu.de, marion.wroblewitz@hhu.de

Abstract

In this paper, we describe our contribution to the SemEval-2023 Task 10 (Subtask A), a shared task on detecting and predicting sexist language. The dataset consists of labeled sexist and non-sexist data targeted towards women acquired from both Reddit and Gab. We present and compare several approaches we experimented with and our final submitted model. Additional error analysis is given to recognize challenges we dealt with in our process. A total of 84 teams participated. Our model ranks 55th overall in Subtask A of the shared task.

1 Introduction

Online sexism is a consistently growing problem on the internet especially on social media platforms such as Reddit and Twitter. When women are the target, it creates a hostile environment and perpetuates sexist beliefs and social stigmas (Dehingia, 2020). Investigating the data gathered from these platforms to not only specifically detect sexist content and certain patterns, but rather explain why it is sexist, gives us a better understanding of the impact it has on women, enhances its interpretability, and helps create a much more welcoming environment.

This paper describes our group's approach for SemEval-2023 Task 10-A (Kirk et al., 2023), which deals with binary sexism classification of sexist and non-sexist content. In order to solve this problem, we worked on fine-tuning transformer language models and on using the provided and additional non-labelled data for an semi-supervised extension of the training data set. Additionally, we compared our results with traditional machine learning classification methods such as a Random Forest or k-Nearest Neighbours classifier. In view of the fact that the provided data had significantly less sexist entries and recognizing the conflict between sexist data and data that is clearly offensive yet non-sexist, we also experimented with data augmentation to overcome the unbalance in the data.

It turns out that our best performing model is a BERT model that is fine-tuned with the manually labeled training data as well as with the automatically labeled additional data that was provided by the organizers. We present the task and background in section 3, our approaches in section 4 and our error analysis in section 5, respectively.

2 Related Work

The detection of hate speech on online platforms has been gaining popularity within the Natural Language Processing community in recent years, focusing on tasks in sub-areas such as identifying racist or abusive content (for literature reviews see Poletto et al., 2021; Jahan and Oussalah, 2021). However, fewer studies have concentrated on hate speech specifically targeted at women (e.g., Jha and Mamidi, 2017).

Classical machine learning techniques based on manual feature engineering use a range of features for hate speech detection (for an overview see Schmidt and Wiegand, 2017), of which word and character n-grams belong to the most indicative ones (Waseem and Hovy, 2016). Robinson et al. (2018) found that automatic feature selection outperforms models with carefully engineered features. Using deep learning ensembles, Zimmerman et al. (2018) report an improvement in F-measure of nearly 5 points compared to classical feature based techniques. Leaderboards from recent shared tasks in offensive speech detection show that transformer architectures perform best (Zampieri et al., 2020). However, Arango et al. (2022) found that the current state-of-the-art methods for hate speech detection have overestimated performance claims due to methodological issues such as data overfitting and sampling issues.

Nowadays, a variety of datasets is available to train hate speech detection models on (for an overview see Poletto et al., 2021; Jahan and Oussalah, 2021). However, as Waseem (2016) and

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1476–1482

Larimore et al. (2021) point out for racist language, annotations may be influenced by annotators racist stereotypes.

In recent times, there has been an increase in research aimed at gaining a deeper understanding of various characteristics of sexism. It is particularly important to distinguish between the various different types and forms of sexism in hate speech such as misogyny, hostile or benevolent sexism (Jha and Mamidi, 2017), subtle or overt sexism, positive sounding stereotypes etc. and how it manifests itself on social media to help identify and classify sexist content (Butt et al., 2021) as well as acknowledge the impact and hurt it causes towards women (Dehingia, 2020). The SemEval-2023 Task 10 tackles this problem (Kirk et al., 2023).

3 Background

SemEval-2023 Task 10 provides participants with a dataset labelled by trained annotators and experts consisting of a total of 20,000 entries,10,000 of which are extracted from the social networking platforms Gab[1] and the remaining 10,000 from Reddit[2] (Kirk et al., 2023). Gab is especially well-known for not restricting users' content by moderation and for attracting users from the 'alt-right'- or the 'incel'-movement for example that express their misogynistic attitude in the form of hate speech (Zannettou et al., 2018; Rieger et al., 2021).

The training data is comprised of 14,000 entries where each entry is separated into columns describing a unique identifier for each entry as "rewire-id", the input text as "text", the labels for Subtask A ('Binary Sexism Detection') as "label-sexist", the labels for Subtask B ('Category of Sexism') as "label-category" and lastly the labels for Subtask C ('Fine-grained Vector of Sexism') as "label-vector". For the development phase 2,000 entries are provided the labels of which have been made available before the final test phase. In the final test phase 4,000 entries are to be labeled by the submitting teams. The distribution of the labels 'sexist' and 'not sexist' in the three data sets ('train', 'dev', 'test') is comparable. About 24% of the entries is labeled 'sexist'. Additional data from Gab and Reddit with 1,000,000 entries each is provided to encourage further training and innovative techniques.

The goal of Subtask A is to build a system that classifies and predicts whether an entry is sexist or not sexist. The results are then used for Subtask B and C that delves into each individual sexist classified entry in depth and categorizes as well as describes the sexist content in a fine-grained manner. This helps thoroughly understand the content and the harm it causes women at such a grand scale.

4 Approaches experimented with

We tackle the classification problem by (a) fine-tuning a pretrained language model (Sec. 4.1) and by (b) classical machine learning methods with feature extraction (Sec. 4.2).[3] The training data provided by the task organizers is split into 1,400 held-out test entries and 12,600 actual training entries for all experiments described in the current section. Splitting occurs as a constant over all experiments.

4.1 Transfer learning

4.1.1 Selecting a baseline model

In order to select a good baseline model for further training, we fine-tuned non-task-specific pretrained BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019) models with our training data (n=12,600). Starting with the idea that sexist speech has a negative sentiment, we also included two BERT models that have already been fine-tuned on a sentiment classification task. We implemented this in order to test the hypothesis that transfer-learning benefits from a model that is pre-trained on sentiment analysis if the task is to detect sexist speech. We chose the model 'distilbert-base-uncased-finetuned-sst-2-english' (HF Canonical Model Maintainers, 2022) which was fine-tuned on the Stanford Sentiment Treebank (Socher et al., 2013) and Twitter-roBERTa-base which was fine-tuned on 58M tweets (Barbieri et al., 2020).

All models were extracted from the HuggingFace platform.[4] Every model was tested on one, two and three epochs respectively. Results for the tested models along with the best epoch settings for each model are shown in Table 1. Only for the bert-base-cased model, we list an additional epoch setting. Although bert-base-cased showed better results when trained on two epochs, we used the bert-base-cased model trained on three epochs for our semi-supervised learning approach. See Section 3.1.3 for more details.

[1] https://gab.com/
[2] https://www.reddit.com/
[3] Our source code is available on GitHub. https://github.com/WiebkePetersen/hhuEDOS2023
[4] https://huggingface.co/

model	acc	f1	e
bert-base-uncased	0.865	0.822	2
bert-base-cased	0.871	0.829	2
bert-base-cased	0.869	0.828	3
roberta-base	0.868	**0.830**	2
distilbert-base-uncased	0.859	0.812	2
" -finetuned-sst-2-english	0.850	0.794	2
twitter-roberta-base-sentiment	0.857	0.818	2

Table 1: Learning results for baseline models. acc = accuracy, f1 = Macro F1 score, e = number of epochs

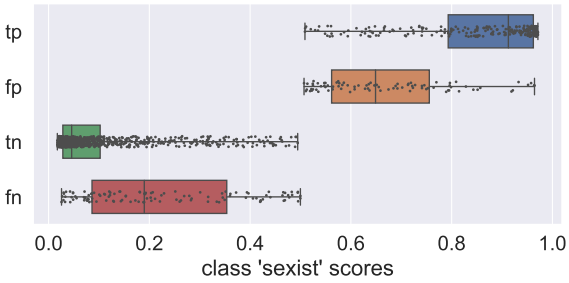

Figure 1: Boxplot of baseline model's softmax scores for class 'sexist' on our own test set (n=1,400).

Due to a lack of GPU computing power, we were unable to test larger BERT models. We came to the conclusion that using a model that was already fine-tuned on a sentiment classification task did not improve the Macro F1 score (see Table 1).

4.1.2 Accounting for imbalanced classes

Given the fact that the provided data is unequally distributed in terms of sexist content, we theorized that augmenting the data by balancing both classes might produce better results.

The training data consists of 3,398 'sexist' and 10,602 'not sexists' entries. To account for this imbalance, we investigated whether simply over- or under-sampling the data improves the results. However, a quick experiment with DistilBERT showed that by under-sampling the majority class 'not sexist' the Macro F1 score declined from 0.82 to 0.77 and by over-sampling the minority class 'sexist' it declined slightly to 0.81. As the final test data in the Shared Task was expected to be as imbalanced as the training data, the best strategy was to continue utilizing the imbalanced training data.

4.1.3 Semi-supervised Learning

As a way of gaining more training data, we used additional unlabeled datasets provided by the task organizers consisting of 2 million entries in total (1 million Gab and 1 million Reddit entries). Based on the experiments in 4.1.1 RoBERTa turned out to be the best performing baseline model. However, due to lack of GPU computation power, we decided to proceed with the cased BERT base model (bert-base-cased) fine-tuned on the training data that only performed slightly worse, yet allowed for shorter training times. We determined to supply the model with 3 epochs instead of 2 for two reasons. First, a few trials with alternative train/test-splits indicated that the model's performance is not consistently better if trained for only 2 epochs. Second, since we use the model to label data for further training, we want to ensure that the model is fitted well on the gold standard training data before we apply unlabeled data on our model.

We applied the model pre-trained on the training data (3 epochs) to the unlabeled Gab and Reddit datasets. Our approach was to use the unlabeled data's softmax score as an indicator of the confidence of the model of the assigned label. Furthermore, we only kept the class labels for which the model confidence exceeded a certain threshold. To determine adequate thresholds, the softmax scores for the 1,400 test entries had to be investigated.

Figure 1 displays a boxplot of the softmax values for class 'sexist' for each of the four cases true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). It appears that only a few entries that have a high softmax value are falsely labeled as 'sexist', yet the false negatives often possess a pretty low softmax value indicating model confidence in labeling them as 'not sexist'. To avoid false negatives, we chose a stricter threshold r for the 'not sexist' class (r=0.015) than the 'sexist' class (r=0.95). The second reason for the decision to use a laxer threshold for the 'sexist' class is that with a stricter threshold our pretrained baseline model would label less entries as 'sexist' and we wouldn't gain sufficient additional training data. Entries with a softmax score of less than 0.015 were categorized as 'not sexist', while entries with a score exceeding 0.95 were classified as 'sexist'. Any other entries that fell within these thresholds were excluded from the training process.

Using these thresholds, we labeled 667,351 Reddit and 864,068 Gab entries as 'not sexist' and 35,803 Reddit and 10,307 Gab entries as 'sexist'. We randomly sampled entries within the newly labeled data to extend our training data while retaining the original class label distribution.

Lastly, we further finetuned our BERT base model with 1 training epoch on the augmented dataset which now consists of 154,883 'not sexist' and 49,145 'sexist' entries. Applying this model to the test data yielded an accuracy score of 0.871 and a Macro F1 score of 0.831. Additionally, we tested whether results improve if the original class label distribution is not retained and equally many examples of both classes are added to the original training set. However, with this configuration, the Macro F1 score dropped to 0.824 (accuracy: 0.867).

4.2 Traditional Machine Learning Classifier

The training data consists of relatively short posts written in informal language using slang vocabulary and non-standard grammar that differ from the texts most transformer models have been pretrained on. Considering that the training corpus is also rather small to effectively fine-tune a transformer model, we experimented with some traditional machine learning classification methods.

For feature extraction we used TF-IDF (Term Frequency – Inverse Document Frequency) that takes into account the frequency of a term within a document (TF) and offsets it by the frequency of the same term across all documents in the corpus (IDF) (Salton and Buckley, 1988). The idea is that TF-IDF can help identifying important features (words or phrases) that are highly indicative of sexist speech by giving addtional weight to words that are more common in sexist posts rather than non-sexists ones.

After preprocessing the training data (removing stop words, punctuation, emojis, special characters, ...), we used the *TfidfVectorizer* function with default parameters from the *Scikit-learn* library for feature extraction (Pedregosa et al., 2011). Different traditional machine learning classifiers from the Scikit-learn classfier library were fitted on the resulting TF-IDF matrix. We experimented with *Decision Tree*, *Random Forest (RF)*, *Ada Boost Classifier*, *Logistic Regression*, and *Multi-layer Perceptron (MLP)* with different parameter settings. The performance of none of these classifiers was satisfying. For an overview of the results see Table 2.

4.3 Model submitted

The submitted model, which performed with a Macro F1 score of 81.85% in the competition, was created as follows. First, we trained a BERT base cased model with 3 epochs on 90% of the training data (see 4.1.1). This model was then used to label additional training data from the supplementary data provided (see 4.1.3). Subsequently, one epoch were trained with the additionally gained training data. Finally, two epochs each were trained with the 10% held-out data and the development data. Table 3 shows the confusion matrix of the predictions made by our submitted model on the tasks test data.

model	**f1**
LogisticRegression	0.68
Decision Tree	0.71
Random Forest, n=25	0.71
Random Forest, n=50	0.71
Random Forest, n=100	0.71
Random Forest, n=150	0.70
AdaBoost, n=50	0.73
AdaBoost, n=100	0.74
AdaBoost, n=250	0.74
AdaBoost, n=500	**0.75**
AdaBoost, n=1000	**0.75**
MLP, (20,10,5)	0.72
MLP, (50,25,10,5)	0.71

Table 2: Results of some traditional machine learning classifiers with different parameter settings. f1 = Macro F1 score.

tn: 2771	fp: 259
fn: 272	tp: 698

Table 3: Confusion matrix of predictions by submitted model on the task test data

5 Error Analysis and conclusion

We conduct a brief analysis of our submitted model on the test set. To get a better understanding of how 'confident' our model is, when it makes predictions, Figure 2 plots the model's softmax scores for the test data.

From Figure 2 we can observe that the false positives have an on average lower prediction score for 'sexist' entries than the true positives. Similarly, the false negatives have a higher prediction score for 'sexist' entries compared to the true negatives. However, Figure 2 also shows that there is no clear boundary line between the true/false positives and the true/false negatives.

To gain a better understanding of our models' weaknesses, we analyzed our results using the more

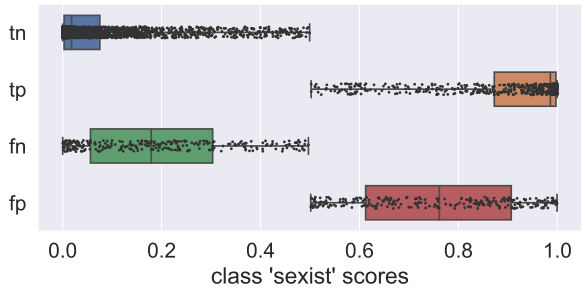

Figure 2: Boxplot of softmax scores for class 'sexist' of submitted model. From top to bottom: tn (true negatives), tp (true positives), fn (false negatives), fp (false positives)

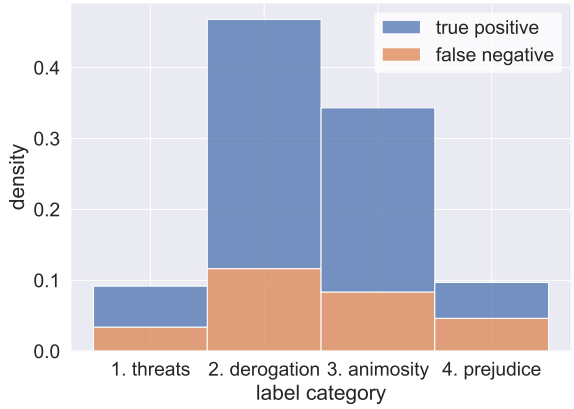

Figure 3: Distribution of false negatives and true positives over the four classes of Subtask B

fine-grained classes from Task B. Figure 3 provides an overview of the data falsely labeled as 'not sexist' (false negatives) and how it is distributed among the four classes from Task B. We compare it with the data that was correctly labeled as 'sexist' (true positives). We found that data labeled as '1. threats' and '4. prejudiced discussions', were often misclassified as 'not sexist'. Both form the two smallest classes, yet only roughly half of the sexist data was classified correctly.

By a manual inspection, we have observed that our model frequently makes incorrect predictions when presented with data containing non-sexist but offensive language. The opposite situation also occurs, when the model falsely predicts a data point/text as 'not sexist' where sexist but non-overtly offensive language is used. The ability of the model to detect sexist content appears to be impacted by the lack or presence of overtly offensive language e.g. use of slurs. The following two examples from the data exemplify this behavior of our model. The first example is a false negative with a softmax score of 0.48, that lacks sexist vocabulary:

> A slow transition starting w/ feminism, false rape reports on campus and the complete indifference in society for the suffering of those men which led me to the men's rights movement, than MGTOW youtubers and TRP.

And the second example is a false positive with a softmax score of 0.94, that uses offensive but not explicitly sexist vocabulary:

> You seem to care enough to threaten him. You are a degenerate shabbos cock sucker who puts a group of people that hates you simply for existing over your white brothers and sisters. You are a disgrace.

As our model performs worse on the 'sexist class' (F1 score: 0.72) compared to the 'not sexist' class (F1 score: 0.91) and as the recall for the 'sexists' class differs significantly between subclasses defined for Subtask B (see Figure 3) it could be worth to try to combine a classifier for Subtask B with one for the current Subtask A.

Acknowledgements

This work evolved from a bachelor seminar at Heinrich-Heine-Universität Düsseldorf. Valuable comments on the work have been given by students from the seminar.

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F Gelbukh. 2021. Sexism identification using bert and data augmentation-exist2021. In *IberLEF@ SEPLN*, pages 381–389.

Nabamallika Dehingia. 2020. When social media is sexist: A call to action against online gender-based violence.

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.

HF Canonical Model Maintainers. 2022. distilbert-base-uncased-finetuned-sst-2-english (revision bfdd146).

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of Hate Speech automatic detection using Natural Language Processing. *arXiv e-prints*, page arXiv:2106.00742.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. 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. *CoRR*, abs/1907.11692.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and George Louis Groh. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and reddit. *Social Media + Society*, 7.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on twitter: Feature engineering v.s. feature selection. In *The Semantic Web: ESWC 2018 Satellite Events*, pages 46–49, Cham. Springer International Publishing.

Gerard Salton and Chris Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):323–328.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1007–1014, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources*

and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).