Performance Comparison of Transformer-Based Models on Twitter Health Mention Classification

Pervaiz Iqbal Khan, Imran Razzak, *Senior Member, IEEE*, Andreas Dengel, and Sheraz Ahmed

***Abstract*— Health mention classification classifies a given piece of text as a health mention or not. However, figurative usage of disease words makes the classification task challenging. To address this challenge, consideration of emojis and surrounding words of the disease names in the text can be helpful. Transformer-based methods are better at capturing the meaning of a word based on its surrounding words compared to traditional methods. However, there are numerous transformer-based methods available and pretrained on natural language processing (NLP) data that are inherently different from Twitter data. Moreover, the size of these models varies in terms of the number of parameters. Hence, it is challenging to decide and choose one of these methods for fine-tuning it on the downstream tasks such as tweet classification. In this work, we experiment with nine widely used transformer methods and compare their performance on the personal health mention classification of tweet data. Furthermore, we analyze the impact of model size on the classification task and provide a brief interpretation of the classification decision made by the best performing classifier. Experimental results show that RoBERTa outperforms all other models by achieving an F1 score of 93%, while two other models perform similarly by achieving an F1 score of 92.5%.**

***Index Terms*— Health mention classification, public health surveillance (PHS), tweet classification.**

I. Introduction

PUBLIC health surveillance (PHS) deals with the collection, analysis, and interpretation of data related to health [1]. The PHS systems aim to detect emergencies, such as pandemics that help authorities in taking preventive actions [2]. The data collection process to train such systems usually involves the crawling of social media platforms such as Twitter, Facebook, and Reddit based on keywords. These keywords can be the names of diseases such as fever, cancer, and depression. However, a keyword-based search may not always retrieve the correct data. Consider three tweets shown in Fig. 1. The tweet in Fig. 1(a) mentions the disease words and indicates the presence of the diseases, but the tweet shown in Fig. 1(b) does not indicate the presence of the disease itself, although it contains the disease word "cough." Consider another tweet example shown in Fig. 1(c), which contains the disease word of "depression" but uses figuratively. The usage of disease words figuratively and in nonhealth fashion poses additional challenges for PHS systems. To address these challenges, one approach is to consider the emojis and the surrounding words of the disease word that give the idea of whether a disease word is used figuratively or not. For example, the tweet shown in Fig. 2(a) contains the disease word "heart attack," and however, the emoji in the tweet indicates that this is a figurative mention. On the other hand, the tweet in Fig. 2(b) containing the emoji of broken heart indicates the presence of Alzheimer's. The tweet shown in Fig. 3 depicts the impact of surrounding words on detecting the figurative mentions of diseases. This tweet contains the disease word "cancer," and surrounding words, such as "diagnosed" and "stage 4," indicate the usage of disease word as health mention.

Existing work uses noncontextual and contextual approaches to learn representations for the text data to classify it as health or nonhealth mention. Jiang *et al.* [3] used the pretrained noncontextual word representations method and passed those representations to long short-term memory networks (LSTMs) [4] and, then, the final classification layer to classify it. Karisani and Agichtein [5] distorted the original noncontextual embedding space that introduced the generalization capabilities to the model. Iyer *et al.* [6] also used the noncontextual embedding and passed them to convolutional neural network (CNN) to classify the tweets. Biddle *et al.* [7] used contextual word representations and combined the various layers of ELMO [8] and BERT [9] and passed the combined embedding to the Bi-LSTM [10].

Transformer-based models are popular methods to solve various natural language processing (NLP) tasks, such as text classification, question answering (QA), and text summarizing [11]. These methods learn the representation of a word in the sentence, considering the surrounding words in a given sentence. However, today, numerous transformer-based models are available that are pretrained on NLP data. Moreover, the size of these models varies in terms of the number of model parameters. Generally, it is believed that the performance on the task improves with an increase in model size [12]. Hence, it is a challenging decision to select one of these models for the downstream task such as classifying a tweet as a health mention or not. The inherent difference of Twitter language style from traditional NLP data further increases complexity. Hence, it seems interesting to compare the performance of various

Manuscript received 24 September 2021; revised 8 December 2021; accepted 30 December 2021. Date of publication 17 February 2022; date of current version 31 May 2023. *(Corresponding author: Pervaiz Iqbal Khan.)*

Pervaiz Iqbal Khan and Andreas Dengel are with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, and also with the Department of Computer Science, TU Kaiserslautern, 67663 Kaiserslautern, Germany (e-mail: pervaiz.khan@dfki.de; andreas.dengel@dfki.de).

Imran Razzak is with the School of Information Technology, Deakin University, Geelong, VIC 2600, Australia (e-mail: imran.razzak@deakin.edu.au).

Sheraz Ahmed is with the German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany (e-mail: sheraz.ahmed@dfki.de).

Digital Object Identifier 10.1109/TCSS.2022.3143768

Fig. 1. PHS Tweet examples. (a) Health mention. (b) Nonhealth mention. (c) Figurative mention.

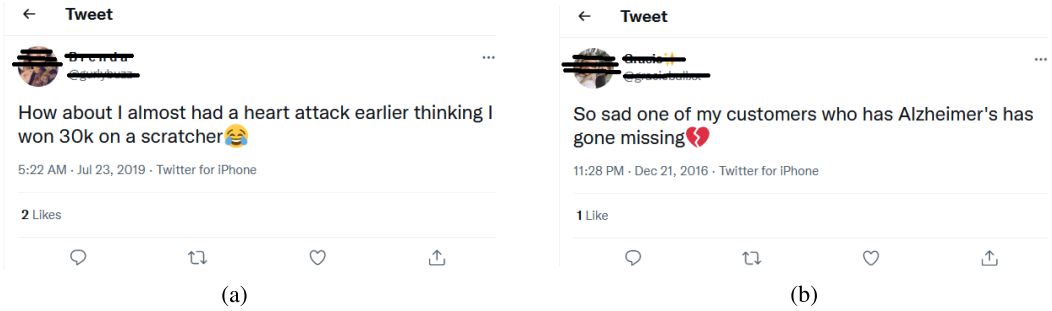

Fig. 2. Emojis are helpful in health classification. (a) Figurative mention. (b) Health mention.

Fig. 3. Context of the disease word is important.

transformer-based word representation methods on the Twitter dataset on the task of health mention classification. In this work, we do a performance comparison of transformer-based models and discuss the impact of model size on classification results. To this end, we experiment with nine of these models belonging to different families and compare their performance in terms of precision, recall, and F1 score.

Generally, deep learning models are considered as black boxes and the factors influencing the decision of the model are unknown. Therefore, we interpret the decision made by the best performing model by taking a few examples from the test set and finding the words that influence the model for making a prediction. The key contributions of this work are given as follows:

1) a comprehensive performance comparison of the nine transformer-based models for health mention classification task on Twitter data;
2) analysis of the impact of the number of model parameters on the performance of health mention classification;
3) visualization of the importance of words contributing toward classification decisions for the best performing model.

The structure of the rest of this article is as follows. Section II provides the overview of related work. Section III presents a brief introduction of transformer-based models used for the comparison. Sections IV and V present experimental details, and results and discussion, respectively. In Section VI, a brief error analysis is provided, whereas in Section VII, words influencing the decision of the classifier are visualized. Section VIII provides the conclusion of this article.

II. Related Work

Pretrained word representations have been extensively used for text classification. Recent NLP methods rely on pretrained word representations [13]–[15]. Initial methods were based on noncontextual statistics such as word2vec [16], and GloVe [17] i.e., the individual word was used as a single representation regardless of its context in the text [8], [14]. For example, the word "Parkinson's" would always be represented with the same vector in each tweet that mentions the disease words "Parkinson's." Recently, the noncontextual approach has been replaced by contextual word representation using transformers. Contextual word representation is a different approach from noncontextual representation and depends on surrounding words while learning representation for a given word. Along with words representation, usage of emojis in the tweets also plays an important role in tasks, such as health mention classification and sentiment analysis. In this section, we present the existing work in the domain of health mention classification and sentiment analysis on the Twitter data using various word representations methods and emojis.

A. Twitter Health Mention Classification

To detect health mention in Twitter data, Karisani and Agichtein [5] presented a new method called word embedding space partitioning and distortion (WESPAD). WESPAD first learned to partition and then distort word representations, which added the generalizing capabilities in the model. Moreover, this method enabled the classifier to require little training examples containing positive health mentions. Despite the improvement in the classifier results and the requirement of little training data, distorting the original word embedding caused information loss. Jiang *et al.* [3] used an LSTM based-classifier to classify a given tweet as a health mention or nonhealth mention. They applied the general preprocessing on the input text and then extracted noncontextual word representations from this text that worked as a classifier input. LSTM-based classifier outperformed the other classifiers, such as support vector machines (SVMs), K-nearest neighbor (KNN), and decision trees. Iyer *et al.* [6] classified the input health-related text as figurative mention or health mention. They used a two-step approach: In the first step, they detected whether the disease word was mentioned in a literal or figurative sense. In the second step, they passed this information as an additional feature along with other features where a CNN-based classifier classified the input text as health mention or not. This additional information from the first step improved the performance of the classifier. Although their method improved the performance, it did not accurately classify figurative mentions, especially the heart attack, one of the most common figurative mentioned words.

To detect figurative mentions of the disease words in a given tweet, Biddle *et al.* [7] used the work of [6] as a baseline. They extended the existing health mention dataset by adding 14k new tweets and four new disease words. They converted the emojis into string representations using the Python library[1] and normalized the URL, as well as the names of the users, mentioned in the Tweet as a preprocessing step. To represent the input text as numeric vector representations, they experimented with both noncontextual representations such as word2vec [16] as well as with contextual representations like ELMO [8] and BERT [9]. They also incorporated sentiment information in the model from the tweet text. For this purpose, they experimented with various sentiment extracting methods, such as WordNet [18], VAD [19], and ULMFit [20]. Bi-LSTM [10] was used as a classifier that took the word representation as an input. The sentiment information was passed as a separate input. The output of the Bi-LSTM was concatenated with the sentiment information to produce a final binary output. Results showed contextual word representation using BERT and VAD as a sentiment extractor gained a performance boost compared to other methods. They validated the performance of the method using tenfold cross validation. An error analysis showed that the misclassifications were caused due to incongruity and simple linguistic patterns. Although using contextual word representation with the sentiment information improved the classification results, this method could not capture the full context of the disease words used.

B. Using Emojis in Tweet Classification

Xu *et al.* [21] studied the effect of emojis on mourning tweet classification. They built models using: 1) only words (OWF) and 2) words and emojis (WEF). To incorporate emojis, they added a token to the tokenizer dictionary for each emoji and then updated the embedding matrix to include each emoji and finally performed fine-tuning. They showed that the use of emojis in the tweets improved the classification results; de Barros *et al.* [22] improved sentiment classification results on two datasets TweetSentBR (TTsBR) [23] and 2000-tweets-BR [24] using emojis with the text. They extracted the emojis from the tweets and processed them separately through transformers and combined both text and emojis representations before classification. Li *et al.* [25] performed multiclass sentiment classification using emojis. To extract emojis characters, they converted text into unicode representation and applied regular expressions to extract unicode related to emojis. Then, they obtained tweet sentiment for each tweet by adding a score for each emoji. They used other features such as linguistic features, sentiment lexicon features, and microblogging features with sentiment scores to perform classification. Biddle *et al.* [7] incorporated sentiment information in the model from the tweet text. The sentiment information was passed as a separate input to sentiment extractors. The output of the sentiment extractor was used as an additional feature to the classifier.

[1] https://pypi.org/project/emoji/

Today, various transformer-based pretrained models are available for learning word representations. Moreover, the size of these models varies, which influences the computational cost and the performance of the downstream task. Hence, it is a challenging task to choose one of these methods for the downstream task. In this work, we selected nine widely used transformer-based models and fine-tuned them on the health mention classification task. We compared the performance of these methods using widely used evaluation metrics such as precision, recall, and F1 score for the classification task.

III. Transformer Models

Transformer models have revolutionized many areas of NLP, such as text classification, QA, text summarizing, and language understanding. BERT [9] is a bidirectional transformer model that is pretrained over large unlabeled text corpus for language understanding and can be fine-tuned for various downstream NLP tasks. It achieves language understanding by using the masked language model (MLM) and next sentence prediction (NSP). BioBERT [26] is a first domain-specific pretrained language model for the biomedical data. Its uses the architecture of BERT. BioBERT is initialized with weights from BERT and then fined-tuned on PubMed abstracts and PMC full-text articles. BioBERT improves the performance on three biomedical text mining tasks of named entity recognition (NER), relation extraction (RE), and QA. XLNet [27] is another transformer-based language model that achieves language understanding by predicting tokens in a given sequence by random order instead of MLM and NSP. This random order-based language modeling helps the model to capture a better relationship between tokens in a given sequence. ALBERT [28] is a lite BERT that reduces the parameters of BERT by using factorized embedding parameterization and cross-layer parameter sharing that allows the model to share the parameters across the layers. Compared to BERT, ALBERT achieves 1.7× faster training and 18× fewer parameters with comparable performance. RoBERTa [29] involves the retraining of BERT with an improved training method and 1000% more data and computation. It removes the NSP task and introduces the dynamic masking of tokens during the training process. Larger training batch size is also a reason for the improvement in the performance of the model on various GLUE benchmark [30] results. BERTweet [31] is another transformer-based model that is pretrained on English Tweets using the architecture of BERT and pretraining procedure of RoBERTa. It outperforms other methods on NER, part-of-speech (POS) tagging, and text classification. The DeBERTa [32] language model improves the previous state-of-the-art models using disentangled attention and enhanced mask decoder. Disentangled attention uses the two separate vectors for each input token that encode its content and positions separately unlike BERT where positional embedding is combined with the word embedding. This allows the model to compute attention weights based on the content as well as on the position of the token. Enhanced mask decoder allows the model to additionally use the absolute position of the token for MLM prediction along with position-based attention. ELECTRA [33] is another transformer-based model that corrupts the token in a sequence by using an alternate token generated by the generator. Then, the discriminative model predicts whether each token in the corrupted input is replaced by a generator or not. It performs comparably to RoBERTa and XLNet while requiring 1/4 of their computing. Open AI GPT-2 [34] is a large transformer-based language model having 1.5 billion parameters and trained on a dataset of 8 million web pages. The training objective of the model is to predict the next words while given all the previous words within the text. GPT-2 is a scale-up of GPT with 10× parameters and 10× training data. The transformer-based models we discussed are based on different pretraining objectives and perform differently on various tasks. These models also vary in the number of parameters. In this work, we fine-tune these models on the downstream task of health mention classification on the Twitter data. We compare the performance of these models in terms of precision, recall, and F1 score while discussing the effect of the number of trainable parameters on the task. Furthermore, we visualize the words that influence the model toward classification decisions.

IV. Experiments

A. Dataset

We performed experiments on the health mention classification dataset of Twitter tweets provided by Biddle *et al.* [7]. The original dataset contained 19 558 tweets, but, at the time of download for experimentation, not all the tweets were available on Twitter that reduced the dataset size from 19 558 to 15 742 tweets. The dataset had a total of 4228, 7322, and 4192 tweets for health mention, nonhealth mention, and figurative mention, respectively. Alzheimer's was the disease with the highest number of tweets with a count of 1715. Heart attack contained the highest number of figurative-mentioned tweets with a total of 1060. Table I shows the diseasewise dataset distribution.

B. Data Preprocessing

To preprocess the data, first of all, we converted all the emojis in the tweets into the text using a python library.[2] For example, the emoji shown in Fig. 2(b) was converted to the text "broken-heart." Then, we removed all the URLs, hashtags, and user mentions from the tweets. We also removed the special characters such as _, ", *, ", -, and: from the tweets.

C. Performance Metrics

To evaluate the performance of the models, we used precision, recall, and F1 score as evaluation metrics. These metrics can be computed as follows.

1) Precision $= \mathrm{TP}/\mathrm{TP} + \mathrm{FP}$.
2) Recall $= \mathrm{TP}/\mathrm{TP} + \mathrm{FN}$.
3) F1 score $= 2 \star (\mathrm{Precision} \star \mathrm{Recall})/\mathrm{Precision} + \mathrm{Recall}$.

Here, the following conditions hold.

1) TP means examples were positive, but the model also predicted them as positives.

[2]https://pypi.org/project/emoji/

TABLE I
DATASET STATISTICS

Disease	Tweet Count	Health Mention	Non-Health Mention	Figurative Mention
Alzheimer's	1,715	249	1,374	92
Cancer	1,691	302	1,239	150
Cough	1,452	331	433	688
Depression	1,579	517	711	351
Fever	1,484	517	342	625
Headache	1,429	791	112	526
Heart attack	1,618	209	349	1,060
Migraine	1,519	904	400	215
Parkinson's	1,568	153	1,362	53
Stroke	1,687	255	1,000	432
Total	15,742	4,228	7,322	4,192

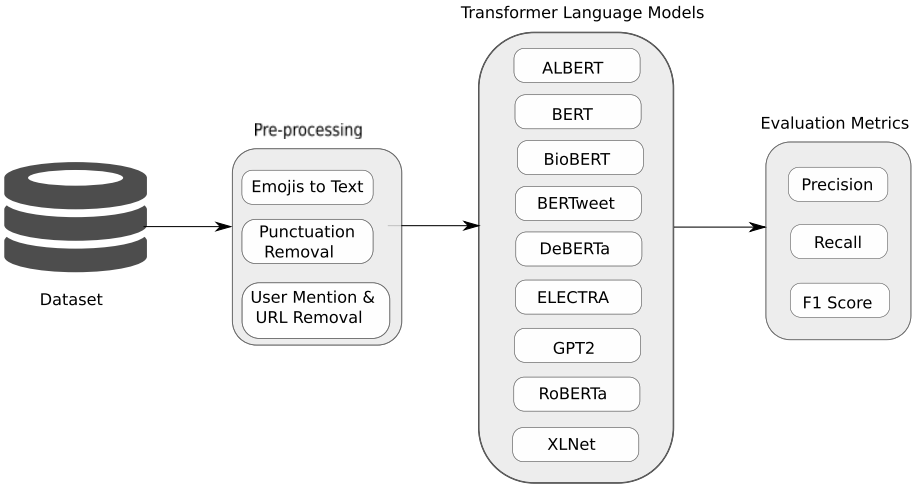

Fig. 4. Experimentation workflow.

2) FP means examples were negative, but the model predicted them as positives.
3) FN means the examples were positive, but the model predicted them as negatives.

D. Experimental Settings

We aimed at comparing the performance of nine transformer-based models on the task of health mention classification for tweets. For this purpose, we selected the uncased version of all the models except for XLNet because only its cased version was available. We used large versions for all the models. First of all, all the tweets went through a preprocessing pipeline that converted emojis in the tweets to a text representation and removed the punctuation, URLs, and user mentions. Then, each tweet was tokenized by the dedicated tokenizer of each model. After that, we fine-tuned transformer models on the health mention classification dataset for classifying the input tweet as health mention or not. Fig. 4 shows the workflow of experiments. We split the dataset into 65%, 15%, and 20% for the train, validation, and test sets, respectively. The maximum sequence length we used was 64. The tweets with lengths greater than 64 were truncated, and the tweets with lengths less than 64 were padded with zeros. We followed the optimization of hyperparameters using cross validation. We experimented with a batch size of 16, 24, 32, 64, and 128 for the BERTweet model. For all the other models, we chose the batch size of 16, 24, and 32 due to computational constraints. We used a fixed learning rate of $1e^{-5}$ and trained models for four epochs. We also used early stopping to prevent overfitting. Then, we picked the best performing models on the validation set and evaluated them on the test set. We used AdamW [35] as an optimization algorithm.

V. RESULTS AND DISCUSSION

Table II summarizes the results of transformer models fine-tuned on the health mention classification dataset. It is evident from the results that BERT and RoBERTa achieved the highest precision of 93%, whereas ELECTRA achieved the lowest precision of 86.5%. On the other hand, BERTweet and RoBERTa achieved the highest recall of 93%. DeBERTa, ELECTRA, and XLNet achieved a recall of 92.5%, 89.5%, and 90%, respectively. GPT-2 was the model with the lowest recall of 87.5%. In terms of F1 score, RoBERTa was the model with the highest performance with 93%, whereas ELECTRA and GPT-2 were the models with the worst performance of 88%. BERTweet and DeBERTa achieved the 92.5% F1 score. BERT achieved the F1 score of 92%, whereas both ALBERT and XLNet achieved the 89% F1 score. Although BioBERT was pretrained on the biomedical text, it achieved the F1 score of 90.5%. This is because Twitter health mention

TABLE II

PRECISION, RECALL, AND F1 SCORE FOR TRANSFORMER-BASED MODELS ON THE TWITTER HEALTH MENTION CLASSIFICATION

Model Name	Precision	Recall	F1 Score
ALBERT [28]	89	88.5	89
BERT [9]	**93**	91	92
BioBERT [26]	92	89	90.5
BERTweet [31]	92	**93**	92.5
DeBERTa [32]	92	92.5	92.5
ELECTRA [33]	86.5	89 .5	88
GPT-2 [34]	89	87.5	88
RoBERTa [29]	**93**	**93**	**93**
XLNet [27]	88.5	90	89

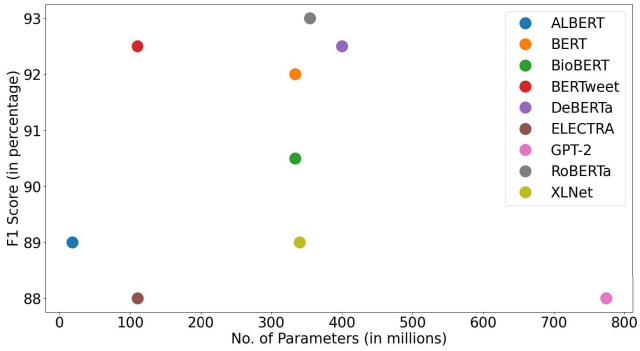

Fig. 5. Plots showing number of parameters versus F1 score for each model.

classification dataset is different from the biomedical dataset used for pretraining the BioBERT.

Table III presents the diseasewise results of each of the models in terms of F1 score. On cancer disease, both BERT and RoBERTa achieved the F1 score of 90%. For cough, BERTweet, DeBERTa, and RoBERTa gained the F1 score of 93.5%. DeBERTa was the best performing model on depression with 86%, whereas BERT, BioBERT, and BERTweet achieved 84% F1 score. On the disease fever, RoBERTa was the best performing model with the F1 score of 94.5%. For headache, BERTweet and RoBERTa achieved 95.5% F1 score. For Parkinson's and stroke diseases, the model RoBERTa was the most successful model with the F1 scores of 85.5% and 96%, respectively. For migraine, DeBERTa was the most successful model with an F1 score of 96%. Diseasewise results show that ALBERT, ELECTRA, GPT-2, and XLNet mostly struggled while making predictions compared to the other models. Fig. 5 shows the number of parameters against the F1 score. Although GPT-2 has the highest number of parameters, it performed worse than ALBERT, BERT, BERTweet, DeBERTa, RoBERTa, and XLNet in terms of F1 score. BERTweet is the model with fewer parameters but performed similar to DeBERTa. ALBERT also has fewer parameters compared to five other models but achieved the F1 scores of 89%. Although BERTweet had fewer parameters than most of the models, it performed a little worse than the best performing model RoBERTa because it was pretrained on the Twitter data. Most of the models could not achieve good classification results because they were pretrained on the NLP text that is different from Twitter language. Although RoBERTa was also pretrained on NLP text, it outperformed other models on the Twitter health mention classification. One reason for this we can think of is the large pretraining data that learn better word representation compared to other models.

Table IV shows the diseasewise precision for each of the models. For Alzheimer's, BioBERT achieved the highest precision of 93%, whereas XLNet achieved the lowest precision of 83.5%. For cancer, BERTweet achieved the highest precision of 90%. BERT outperformed other models for cough, fever, and heart attack by having a precision of 95.5%, 94.5%, and 95.5% respectively. ELECTRA performed the worst on diseases of cancer, cough, depression, fever, and heart attack with a precision of 83%, 84.5%, 78%, 87%, and 76.5%, respectively. GPT-2 achieved the minimum precision of 77% and 87.5% for Parkinson's and headache, respectively. For headache, BERTweet had the highest precision of 96%. RoBERTa outperformed other models on stroke with the precision of 97%, whereas BioBERT was the best model with 89% of precision on Parkinson's. DeBERTa outperformed other models on migraine with 96% precision. For depression, DeBERTa had a precision of 85%.

Diseasewise recall is given in Table V. For Alzheimer's DeBERTa achieved the highest recall of 93%, and BioBERT achieved the lowest recall of 77%. For cancer, BERTweet achieved the highest recall of 93%, and ALBERT achieved the lowest recall of 82.5%. For cough, fever, and Parkinson's, RoBERTa had the highest recall of 95.5%, 95%, and 87.5%, respectively. For headache, BERTweet and RoBERTa performed equally with 95.5% recall, whereas for heart attack, both BERTweet and DeBERTa had the recall of 90.5%. Both XLNet and RoBERTa had the maximum recall of 94.5% for stroke. For cough, ALBERT had the lowest recall of 85%.

For fever, Parkinson's, and stroke, GPT-2 had the lowest recall of 88%, 76%, and 83.5%, respectively, whereas for fever, BioBERT had the lowest recall of 85.5%. For migraine, XLNet achieved the lowest recall of 86.5%.

Embeddings learned by the transformer methods are visualized in Fig. 6. The embeddings for the methods of ALBERT, ELECTRA, GPT-2, and XLNet are cluttered and hence not easily separable for the classifier. On the other hand, embeddings learned by BERT and BERTweet are far away for each of the classes and can be separated easily by the classifier. The embeddings for the model RoBERTa are more compact for both the classes compared to the embeddings of other models, which results in the improved classification performance of the model. BERT learned more separable embeddings for cough, fever, headache, and migraine. BERTweet learned better embeddings for fever and headache, but it struggled for learning better representations of Parkinson's, cancer, and depression. Electra had a tough time while learning the embeddings for heart attack, depression, Parkinson's, and fever. GPT-2 could not learn good representations for depression, heart attack, Parkinson's, and migraine. RoBERTa learned good representations for cough, fever, headache, and heart attack, whereas it struggled for depression. XLNet could not learn good representations for most of the diseases.

The reason for the worst performance of models on Alzheimer's, cancer, and Parkinson's disease is the imbalanced

TABLE III

DISEASEWISE PERFORMANCE MEASURED IN TERMS OF F1 SCORE FOR EACH OF THE MODELS

Model	Alzheimer's	Cancer	Cough	Depression	Fever	Headache	Heart Attack	Migraine	Parkinson's	Stroke
ALBERT [28]	85.5	84.5	86	79	89.5	92	78.5	92.5	80.5	90
BERT [9]	87.5	**90**	92.5	84	93	92.5	**92.5**	93	84	90
BioBERT [26]	83	87	91.5	84	87	92.5	86.5	93	84.5	90.5
BERTweet [31]	88	81.5	**93.5**	84	93.5	**95.5**	91	94	83	90.5
DeBERTa [32]	89.5	89	**93.5**	**86**	92	92.5	80.5	**96**	84	83.5
ELECTRA [33]	84.5	84	86.5	78	88	90	78	92.5	79.5	88.5
GPT-2 [34]	83	87	90	82	89.5	87.5	78.5	89	76.5	86
RoBERTa [29]	**90.5**	**90**	**93.5**	84.5	**94.5**	**95.5**	90	95.5	**85.5**	**96**
XLNet [27]	86.5	88	91	80	91.5	91.5	87	88	80	93

TABLE IV

DISEASEWISE PERFORMANCE MEASURED IN TERMS OF PRECISION FOR EACH OF THE MODELS

Model	Alzheimer's	Cancer	Cough	Depression	Fever	Headache	Heart Attack	Migraine	Parkinson's	Stroke
ALBERT [28]	86	88	87.5	78.5	90.5	92	83.5	92.5	81	92
BERT [9]	92	89	**95.5**	83.5	**94.5**	92.5	**95.5**	93	83.5	94.5
BioBERT [26]	**93**	88	92	84	90.5	92.5	94	92.5	89	93
BERTweet [31]	91.5	**90**	91.5	83	94	**96**	91.5	94	81.5	92.5
DeBERTa [32]	87	88	93	**85**	93.5	92.5	90.5	**96**	82	93
ELECTRA [33]	85	83	84.5	78	87	91	76.5	93	78.5	90
GPT-2 [34]	85.5	88	89.5	83	92	87.5	79.5	89	77	90
RoBERTa [29]	91.5	89	92.5	84	94	95.5	94	95.5	**84**	**97**
XLNet [27]	83.5	86.5	89.5	79.5	91	92	85.5	90.5	78	92

TABLE V

DISEASEWISE PERFORMANCE MEASURED IN TERMS OF RECALL FOR EACH OF THE MODELS

Model	Alzheimer's	Cancer	Cough	Depression	Fever	Headache	Heart Attack	Migraine	Parkinson's	Stroke
ALBERT [28]	84.5	82.5	85	81	89	92	74.5	93	80	98
BERT [9]	84.5	91	90	84.5	92	93	90	94.5	85.5	87.5
BioBERT [26]	77	85.5	91.5	84	85.5	93	81.5	93.5	81	88.5
BERTweet [31]	89	**93**	94.5	84.5	93	**95.5**	**90.5**	94	85	89
DeBERTa [32]	**93**	91	93.5	**88**	90	92.5	**90.5**	**96**	87	93.5
ELECTRA [33]	84.5	85	89	81.5	88.5	89.5	81	92	81.5	86.5
GPT-2 [34]	81	86	91	81	88	88	77.5	89.5	76	83.5
RoBERTa [29]	89	91	**95.5**	85.5	**95**	**95.5**	86.5	**96**	**87.5**	**94.5**
XLNet [27]	91	89	92.5	81.5	91.5	91	88.5	86.5	83	**94.5**

TABLE VI

VISUALIZATIONS SHOWING WORDS THAT ARE FOCUSED BY THE MODEL FOR MAKING A PREDICTION. GREEN HIGHLIGHTED WORDS INFLUENCE THE MODEL TOWARD CLASSIFICATION DECISION, WHILE RED HIGHLIGHTED WORDS INFLUENCE IT AGAINST THE DECISION

Sr.No	Ground Truth	Prediction	Word Importance
1	Health Mention	Health Mention	#s just got diagnosed with depression clown face #/s
2	Health Mention	Health Mention	#s I officially have the worse headache . #/s
3	Non-health Mention	Non-health Mention	#s I almost had a heart attack at the light face with tears of joy #/s
4	Non-health Mention	Non-health Mention	#s Son i had a heart attack face with tears of joy face with tears of joy #/s
5	Non-health Mention	Non-health Mention	#s When someone sends me a text of just my name I have a heart attack #/s
6	Non-health Mention	Health Mention	#s I was quiet on social media for about a week and followers have diagnosed me with depression #/s

data in the dataset. On the other hand, for cough, migraine, and headache, most of the models performed well due to balanced data for health mention and nonhealth mention examples. Stroke also had imbalanced class examples, but RoBERTa surprisingly performed well on it.

VI. ERROR ANALYSIS

In this section, we provide a brief error analysis for the best performing model RoBERTa. We randomly selected some of the misclassified tweets from the test set. Our analysis shows that misclassified tweets fall into one of the four categories: 1) a tweet has an image or a video link that contains further information; 2) a tweet with incorrect labeling; 3) a tweet that is difficult for a human to classify; and 4) a tweet with a little content. Examples of the tweets from the first category are: "The guy in hospital cause of Cancer" and "Re Learning to Trust My Own Body with Parkinson's Disease." A few examples from the second category are: "Heart breaking for family of 6 whose van swept away in flood. Parents had Alzheimer's and were holding hands," and "Supreme Court Justice Ruth Bader Ginsburg tells NPR that despite battling cancer for the third time, she is not going anywhere anytime soon I am very much alive." These examples are labeled as "nonhealth mentions," but actually, these are "health mentions." Now, consider these two tweets: "Depression is kicking my butt right now" and "I died because of my depression." These tweets contain little content and are labeled as nonhealth mentions. However, different humans will not agree on its label, and

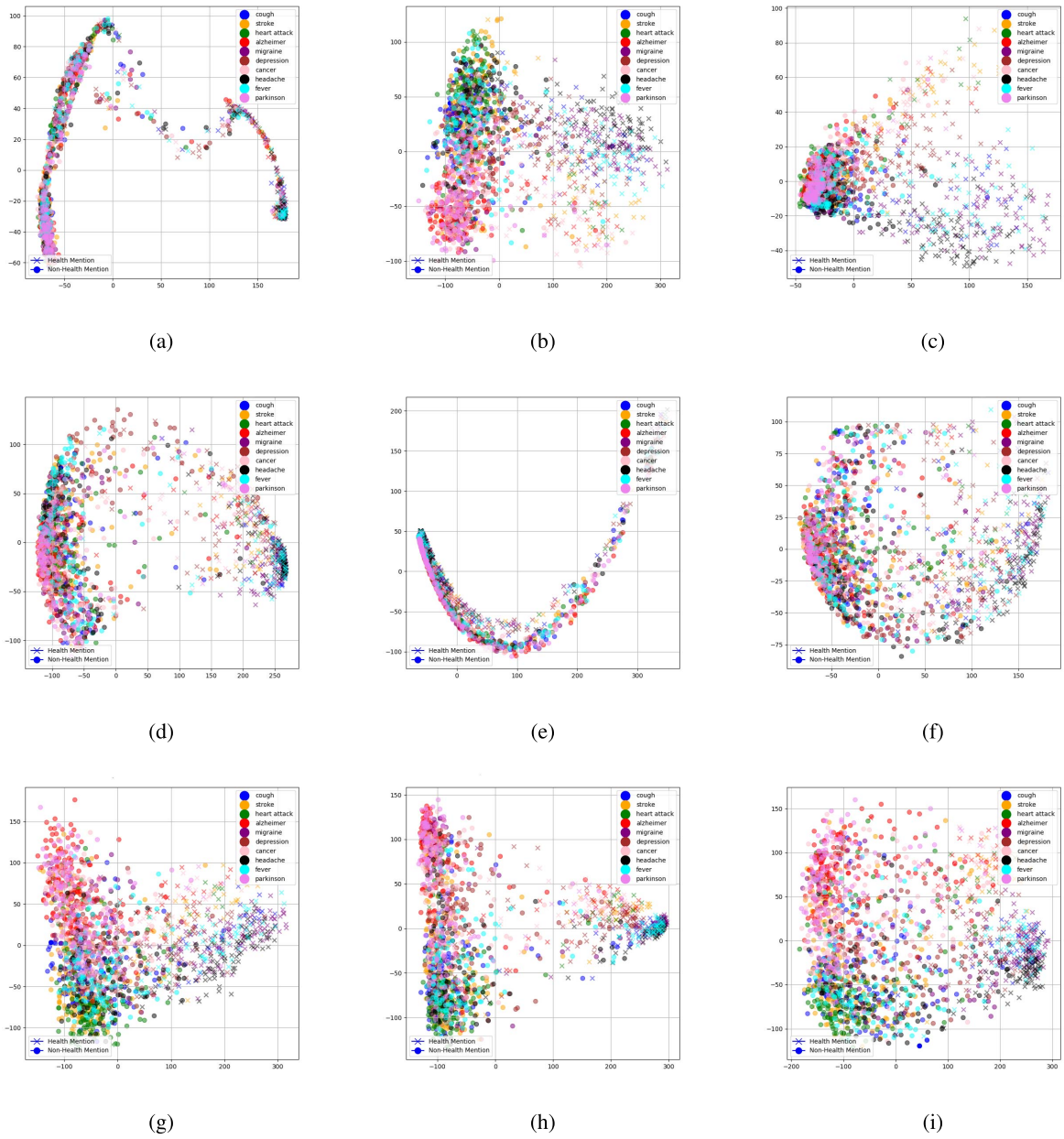

Fig. 6. Embedding of transformer-based models for health mention classification. (a) ALBERT. (b) BERT. (c) BioBERT. (d) BERTweet. (e) DeBERTa. (f) ELECTRA. (g) GPT-2. (h) RoBERTa. (i) XLNet.

hence, these tweets fall into the third category. Examples of tweets with little content are: "Cough not getting any better," "I died because of my depression," and "When you have a headache and the Advil HITS." Some of these examples fall into multiple categories. For example, the tweet "Im fucking depression" has little content, and it is labeled as nonhealth mention but can be categorized as health mention. Another example of such kind of tweet is "I died because of my depression."

VII. Prediction Interpretation

Deep learning models are black box in nature and it is unclear how these models reason. Explainable artificial intelligence (AI) aims to describe the internals of the models in a human-understandable manner [36]. To visualize the words that influence the model toward the classification decision, we used the transformers interpret [37]. We randomly selected a few tweets from the test set and analyzed the predictions made by the best performing model, i.e., RoBERTa. As shown in Table VI, the first tweet example contains a health mention, and the model also classifies this tweet as a health mention. For the classification decision, the words "diagnosed with" and "clown" influence the model toward predicted class, i.e., health mention. In the second example, the word "worse" influences the model for predicting health mention. In the third tweet example, the words, "heart," "light face," and "joy" contribute toward the predicted class, whereas the words "attack" and "with tears" contribute against the predicted

class. This makes sense because the word "joy" indicates the figurative mention, and the word "tears" indicates the presence of disease; however, the overall effect of words results in nonhealth mention prediction. In the fourth and fifth tweet examples, the words "heart," "tears," and "attack" influence the model for health mention prediction, but the words "face" and "text" pull the model toward nonhealth mention prediction. Although the tweet in the sixth example is nonhealth mention, the words "with depression" force the model to predict health mention class. The analysis of these examples shows that the emojis and context of the words play an important role in classification decisions.

VIII. Conclusion

In this article, we compared the performance of various transformer-based models on a Twitter dataset for a personal health classification task. We showed that although models such as XLNet improved the text classification results over BERT in their original paper, this was not the case for health mention classification. Moreover, we analyzed the impact of model size on classifier decisions and provided a brief interpretation of classification decisions made by the classifier. Results show that RoBERTa achieved the highest F1 score of 93%, whereas BERTweet and DeBERTa achieved the F1 score of 92.5%. BERT achieved the precision of 93% but an F1 score of 92%. Experimental results show that the model size did not improve the performance on the downstream task of health mention classification. Based on the results, it can be concluded that if performance is the primary consideration, RoBERTa is the model to be fine-tuned on the task of health mention classification. On the other hand, if the computation cost is the primary consideration, BERTweet is a good option as it requires fewer training parameters with minor performance degradation.

References

[1] WHO. (2021). *Epidemic Intelligence–Systematic Event Detection*. [Online]. Available: https://www.who.int/csr/alertresponse/epidemicintelligence/en/

[2] M. Kogan, L. Palen, and K. M. Anderson, "Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy," in *Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, 2015, pp. 981–993.

[3] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, and G. R. Bernard, "Identifying tweets of personal health experience through word embedding and LSTM neural network," *BMC Bioinf.*, vol. 19, no. S8, p. 210, Jun. 2018.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] P. Karisani and E. Agichtein, "Did you really just have a heart attack? Towards robust detection of personal health mentions in social media," in *Proc. World Wide Web Conf.*, 2018, pp. 137–146.

[6] A. Iyer, A. Joshi, S. Karimi, R. Sparks, and C. Paris, "Figurative usage detection of symptom words to improve personal health mention detection," 2019, *arXiv:1906.05466*.

[7] R. Biddle, A. Joshi, S. Liu, C. Paris, and G. Xu, "Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter," in *Proc. Web Conf.*, Apr. 2020, pp. 1217–1227.

[8] M. E. Peters *et al.*, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[10] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[11] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Springer, 2019, pp. 194–206.

[12] J. Hestness, N. Ardalani, and G. Diamos, "Beyond human-level accuracy: Computational challenges in deep learning," in *Proc. 24th Symp. Princ. Pract. Parallel Program.*, 2019, pp. 1–14.

[13] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? Adapting pretrained representations to diverse tasks," 2019, *arXiv:1903.05987*.

[14] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," 2019, *arXiv:1903.08855*.

[15] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-T. Yih, "Dissecting contextual word embeddings: Architecture and representation," 2018, *arXiv:1808.08949*.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[17] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[18] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. LREC*, vol. 10, 2010, pp. 2200–2204.

[19] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 174–184.

[20] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.

[21] X. Xu, R. Manrique, and B. Pereira Nunes, "RIP emojis and words to contextualize mourning on Twitter," in *Proc. 32st ACM Conf. Hypertext Social Media*, Aug. 2021, pp. 257–263.

[22] T. M. de Barros, H. Pedrini, and Z. Dias, "Leveraging emoji to improve sentiment classification of tweets," in *Proc. 36th Annu. ACM Symp. Appl. Comput.*, Mar. 2021, pp. 845–852.

[23] H. B. Brum and M. das Graças Volpe Nunes, "Building a sentiment corpus of tweets in Brazilian Portuguese," 2017, *arXiv:1712.08917*.

[24] D. Vitório, E. Souza, I. Teles, and A. L. Oliveira, "Investigating opinion mining through language varieties: A case study of Brazilian and European portuguese tweets," in *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Brazil: SBC OpenLib, 2017, pp. 43–52.

[25] M. Li, E. Ch'ng, A. Y. L. Chong, and S. See, "Multi-class Twitter sentiment classification with emojis," *Ind. Manage. Data Syst.*, vol. 118, no. 9, pp. 1804–1820, Sep. 2018.

[26] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinf.*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[27] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.

[28] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[29] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[30] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.

[31] D. Quoc Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," 2020, *arXiv:2005.10200*.

[32] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.

[33] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.

[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[35] N. Landro, I. Gallo, and R. La Grassa, "Mixing Adam and SGD: A combined optimization method," 2020, *arXiv:2011.08042*.

[36] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 80–89.
[37] C. Pierse. (Feb. 2021). *Transformers Interpret*. [Online]. Available: https://github.com/cdpierse/transformers-interpret

Pervaiz Iqbal Khan received the bachelor's degree in computer engineering and the master's degree in computer science from the University of Engineering and Technology, Lahore, Pakistan, in 2007 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer science with the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany, under the supervision of Dr. Andreas Dengel. His research focus lies on improving the healthcare services using artificial intelligence.

Imran Razzak (Senior Member, IEEE) has been a Senior Lecturer of computer science with the School of Information Technology, Deakin University, Geelong, VIC, Australia, since November 2019. He has published more than 120 papers in reputed journals and conferences. He is the author of one book and the inventor of one patent on face recognition. He has attracted a research grant of 1.2 million AUD and has successfully delivered several research projects. He has applied machine learning methods with emphasis on natural language processing and image analysis to solve real-world problems related to health, finance, and social media. His area of interest includes machine learning with its application spans a broad range of topics.

Andreas Dengel received the Diploma degree in computer science from TU Kaiserslautern, Kaiserslautern, Germany, in 1986, and the Ph.D. degree from the University of Stuttgart, Stuttgart, Germany, in 1989.

In 1993, he became a Professor in computer science at TU Kaiserslautern, where he holds the Chair of Knowledge-Based Systems. In 2009, he was appointed as a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University, Sakai, Japan. He also worked at IBM Germany, Mainz, Germany; Siemens, Munich-Perlach, Germany; Xerox Parc, Palo Alto, CA, USA. He is currently the Scientific Director of the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. He is a coeditor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications. He has supervised more than 170 Ph.D. and master theses. His main scientific emphases are in the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media.

Dr. Dengel is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is a fellow of the International Association of Pattern Recognition (IAPR). He received many prominent international awards.

Sheraz Ahmed received the M.S. and Ph.D. degrees in computer science from TU Kaiserslautern, Kaiserslautern, Germany, in 2011 and 2015, respectively, under the supervision of Dr. Andreas Dengel and Dr. Marcus Liwicki. His Ph.D. topic was on generic methods for information segmentation in document images.

He is currently a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, where he is leading the area of time series analysis and life science. Over the last few years, he has primarily worked on the development of various systems for information segmentation in document images. His research interests include document understanding, generic segmentation framework for documents, pattern recognition, anomaly detection, gene analysis, medical image analysis, and natural language processing. He has more than 80 publications on the said and related topics, including three journal articles and two book chapters.

International Journal on Document Analysis and Recognition (IJDAR), International Conference on Document Analysis and Recognition (ICDAR), International Conference on Frontiers in Handwriting Recognition (ICFHR), and Document Analysis Systems (DAS).