Literature Review

**Hielke Muizelaar Master’s Thesis**

## **Full source list**

Amounts of citations are according to Google Scholar.

1. **DocBERT: BERT for Document Classification**

Abstract: We present, to our knowledge, the first application of BERT to document classification. A few characteristics of the task might lead one to think that BERT is not the most appropriate model: syntactic structures matter less for content categories, documents can often be longer than typical BERT input, and documents often have multiple labels. Nevertheless, we show that a straightforward classification model using BERT is able to achieve the state of the art across four popular datasets. To address the computational expense associated with BERT inference, we distill knowledge from BERT-large to small bidirectional LSTMs, reaching BERT-base parity on multiple datasets using 30x fewer parameters. The primary contribution of our paper is improved baselines that can provide the foundation for future work.

Year: 2019

Citations: 299

Full text: Yes

1. **Identification of asthma control factor in clinical notes using a hybrid deep learning model.**

Abstract: BACKGROUND: There are significant variabilities in guideline-concordant documentation in asthma care. However, assessing clinician’s documentation is not feasible using only structured data but requires labor-intensive chart review of electronic health records (EHRs). A certain guideline element in asthma control factors, such as review inhaler techniques, requires context understanding to correctly capture from HER free text. METHODS: The study data consist of two sets: (1) manual chart reviewed data-1039 clinical notes of 300 patients with asthma diagnosis, and (2) weakly abellin data (distant supervision)-27,363 clinical notes from 800 patients with asthma diagnosis. A context-aware language model, Bidirectional Encoder Representations from Transformers (BERT) was developed to identify inhaler techniques in HER free text. Both original BERT and clinical BioBERT (cBERT) were applied with a cost-sensitivity to deal with imbalanced data. The distant supervision using weak labels by rules was also incorporated to augment the training set and alleviate a costly manual abelling process in the development of a deep learning algorithm. A hybrid approach using post-hoc rules was also explored to fix BERT model errors. The performance of BERT with/without distant supervision, hybrid, and rule-based models were compared in precision, recall, F-score, and accuracy. RESULTS: The BERT models on the original data performed similar to a rule-based model in F1-score (0.837, 0.845, and 0.838 for rules, BERT, and cBERT, respectively). The BERT models with distant supervision produced higher performance (0.853 and 0.880 for BERT and cBERT, respectively) than without distant supervision and a rule-based model. The hybrid models performed best in F1-score of 0.877 and 0.904 over the distant supervision on BERT and cBERT. CONCLUSIONS: The proposed BERT models with distant supervision demonstrated its capability to identify inhaler techniques in HER free text, and outperformed both the rule-based model and BERT models trained on the original data. With a distant supervision approach, we may alleviate costly manual chart review to generate the large training data required in most deep learning-based models. A hybrid model was able to fix BERT model errors and further improve the performance.

Year: 2021

Citations: 6

Full text: Yes

1. **Analysis of Language Embeddings for Classification of Unstructured Pathology Reports**

Abstract: A pathology report is one of the most significant medical documents providing interpretive insights into the visual appearance of the patient’s biopsy sample. In digital pathology, high-resolution images of tissue samples are stored along with pathology reports. Despite the valuable information that pathology reports hold, they are not used in any systematic manner to promote computational pathology. In this work, we focus on analyzing the reports, which are generally unstructured documents written in English with sophisticated and highly specialized medical terminology. We provide a comparative analysis of various embedding models like BioBERT, Clinical BioBERT, BioMed-RoBERTa and Term Frequency-Inverse Document Frequency (TF-IDF), a traditional NLP technique, as well as the combination of embeddings from pre-trained models with TF-IDF. Our results demonstrate the effectiveness of various word embedding techniques for pathology reports.

Year: 2021

Citations: 2

Full text: Yes

1. **Publicly Available Clinical BERT Embeddings**

Abstract: Contextual word embedding models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) have dramatically improved performance for many natural language processing (NLP) tasks in recent months. However, these models have been minimally explored on specialty corpora, such as clinical text; moreover, in the clinical domain, no publicly-available pre-trained BERT models yet exist. In this work, we address this need by exploring and releasing BERT models for clinical text: one for generic clinical text and another for discharge summaries specifically. We demonstrate that using a domain-specific model yields performance improvements on three common clinical NLP tasks as compared to nonspecific embeddings. These domain-specific models are not as performant on two clinical de-identification tasks, and argue that this is a natural consequence of the differences between de-identified source text and synthetically non de-identified task text.

Year: 2019

Citations: 1.158

Full text: Yes

1. **MLT-DFKI at CLEF eHealth 2019: Multi-label Classiﬁcation of ICD-10 Codes with BERT**

Abstract: With the adoption of electronic health record (EHR) systems, hospitals and clinical institutes have access to large amounts of heterogeneous patient data. Such data consists of structured (insurance details, billing data, lab results etc.) and unstructured (doctor notes, admission and discharge details, medication steps etc.) documents, of which, latter is of great signiﬁcance to apply natural language processing (NLP) techniques. In parallel, recent advancements in transfer learning for NLP has pushed the state-of-the-art to new limits on many language understanding tasks. Therefore, in this paper, we present team DFKIMLT’s participation at CLEF eHealth 2019 Task 1 of automatically assigning ICD-10 codes to non-technical summaries (NTSs) of animal experiments where we use various architectures in multi-label classiﬁcation setting and demonstrate the eﬀectiveness of transfer learning with pre-trained language representation model BERT (Bidirectional Encoder Representations from Transformers) and its recent variant BioBERT. We ﬁrst translate task documents from German to English using automatic translation system and then use BioBERT which achieves an F1-micro of 73.02% on submitted run as evaluated by the challenge.

Year: 2019

Citations: 50

Full text: Yes

1. **Classification of Medical Image Notes for Image Labeling by Using MinBERT**

Abstract: The lack of labeled image data poses a serious challenge to the application of artificial intelligence (AI) in medical image diagnosis. Medical image notes contain valuable patient information that could be used to label images for machine learning tasks. However, most image note texts are unstructured with heterogeneity and short-paragraph characters, which fail traditional keyword-based techniques. We utilized a deep learning approach to recover missing labels for medical image notes automatically by using a combination of deep word embedding and deep neural network classifiers. Bidirectional encoder representations from transformers trained on medical image notes corpus (MinBERT) were proposed. We applied the proposed techniques to two typical classification tasks: Medical image type identification and clinical diagnosis identification. The two methods significantly outperformed baseline methods and presented high accuracies of 99.56% and 99.72% in image type identification and of 94.56% and 92.45% in clinical diagnosis identification. Visualization analysis further indicated that word embedding could efficiently capture semantic similarities and regularities across diverse expressions. Results indicated that our proposed framework could accurately recover the missing label information of medical images through the automatic extraction of electronic medical record information. Hence, it could serve as a powerful tool for exploring useful training data in various medical AI applications.

Year: 2023

Citations: -

Full text: Yes

1. **Natural Language Processing Model for Identifying Critical Findings-A Multi-Institutional Study.**

Abstract: Improving detection and follow-up of recommendations made in radiology reports is a critical unmet need. The long and unstructured nature of radiology reports limits the ability of clinicians to assimilate the full report and identify all the pertinent information for prioritizing the critical cases. We developed an automated NLP pipeline using a transformer-based ClinicalBERT++ model which was fine-tuned on 3 M radiology reports and compared against the traditional BERT model. We validated the models on both internal hold-out ED cases from EUH as well as external cases from Mayo Clinic. We also evaluated the model by combining different sections of the radiology reports. On the internal test set of 3819 reports, the ClinicalBERT++ model achieved 0.96 f1-score while the BERT also achieved the same performance using the reason for exam and impression sections. However, ClinicalBERT++ outperformed BERT on the external test dataset of 2039 reports and achieved the highest performance for classifying critical finding reports (0.81 precision and 0.54 recall). The ClinicalBERT++ model has been successfully applied to large-scale radiology reports from 5 different sites. Automated NLP system that can analyze free-text radiology reports, along with the reason for the exam, to identify critical radiology findings and recommendations could enable automated alert notifications to clinicians about the need for clinical follow-up. The clinical significance of our proposed model is that it could be used as an additional layer of safeguard to clinical practice and reduce the chance of important findings reported in a radiology report is not overlooked by clinicians as well as provide a way to retrospectively track large hospital databases for evaluating the documentation of the critical findings.

Year: 2022

` Citations: -

1. **BERT-NL a set of language models pre-trained on the Dutch SoNaR corpus**

Abstract: -

Year: 2019

Citations: 2

Full text: No

1. **Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports.**

Abstract: MOTIVATION: The development of deep, bidirectional transformers such as Bidirectional Encoder Representations from Transformers (BERT) led to an outperformance of several Natural Language Processing (NLP) benchmarks. Especially in radiology, large amounts of free-text data are generated in daily clinical workflow. These report texts could be of particular use for the generation of labels in machine learning, especially for image classification. However, as report texts are mostly unstructured, advanced NLP methods are needed to enable accurate text classification. While neural networks can be used for this purpose, they must first be trained on large amounts of manually labelled data to achieve good results. In contrast, BERT models can be pre-trained on unlabelled data and then only require fine tuning on a small amount of manually labelled data to achieve even better results. RESULTS: Using BERT to identify the most important findings in intensive care chest radiograph reports, we achieve areas under the receiver operation characteristics curve of 0.98 for congestion, 0.97 for effusion, 0.97 for consolidation and 0.99 for pneumothorax, surpassing the accuracy of previous approaches with comparatively little annotation effort. Our approach could therefore help to improve information extraction from free-text medical reports. Availability  and implementationWe make the source code for fine-tuning the BERT-models freely available at https://github.com/fast-raidiology/bert-for-radiology. SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

Year: 2020

Citations: 35

Full text: Yes

1. **Emotional RobBERT and Insensitive BERTje: Combining Transformers and Affect Lexica for Dutch Emotion Detection**

Abstract: In a ﬁrst step towards improving Dutch emotion detection, we try to combine the Dutch transformer models BERTje and RobBERT with lexicon-based methods. We propose two architectures: one in which lexicon information is directly injected into the transformer model and a meta-learning approach where predictions from transformers are combined with lexicon features. The models are tested on 1,000 Dutch tweets and 1,000 captions from TV-shows which have been manually annotated with emotion categories and dimensions. We ﬁnd that RobBERT clearly outperforms BERTje, but that directly adding lexicon information to transformers does not improve performance. In the meta-learning approach, lexicon information does have a positive effect on BERTje, but not on RobBERT. This suggests that more emotional information is already contained within this latter language model.

Year: 2021

Citations: 2

Full text: Yes

1. **A Pre-trained Clinical Language Model for Acute Kidney Injury**

Abstract: Pre-trained contextual language models such as BERT have dramatically improved performances for many NLP tasks recently. However, few have explored BERT on specific medical domain tasks such as early prediction for Acute Kidney Injury (AKI). Since much of the clinical information is contained in clinical notes that are largely unstructured text, in this paper, we present an AKI domain-specific pre-trained language model based on BERT (AKI-BERT) that could be used to mine the clinical notes for AKI early prediction. Our experiments on MIMIC-III dataset demonstrate that AKI-BERT can yield performance improvements for AKI early prediction.

Year: 2020

Citations: 3

Full text: Yes

1. **Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing.**

Abstract: Allergic reactions to medication range from mild to severe or even life-threatening. Proper documentation of patient allergy information is critical for safe prescription, avoiding drug interactions, and reducing healthcare costs. Allergy information is regularly obtained during the medical interview, but is often poorly documented in electronic health records (EHRs). While many EHRs allow for structured adverse drug reaction (ADR) reporting, a free-text entry is still common. The resulting information is neither interoperable nor easily reusable for other applications, such as clinical decision support systems and prescription alerts. Current approaches require pharmacists to review and code ADRs documented by healthcare professionals. Recently, the effectiveness of machine algorithms in natural language processing (NLP) has been widely demonstrated. Our study aims to develop and evaluate different NLP algorithms that can encode unstructured ADRs stored in EHRs into institutional symptom terms. Our dataset consists of 79,712 pharmacist-reviewed drug allergy records. We evaluated three NLP techniques: Naive Bayes-Support Vector Machine (NB-SVM), Universal Language Model Fine-tuning (ULMFiT), and Bidirectional Encoder Representations from Transformers (BERT). We tested different general-domain pre-trained BERT models, including mBERT, XLM-RoBERTa, and WanchanBERTa, as well as our domain-specific AllergyRoBERTa, which was pre-trained from scratch on our corpus. Overall, BERT models had the highest performance. NB-SVM outperformed ULMFiT and BERT for several symptom terms that are not frequently coded. The ensemble model achieved an exact match ratio of 95.33%, a F1 score of 98.88%, and a mean average precision of 97.07% for the 36 most frequently coded symptom terms. The model was then further developed into a symptom term suggestion system and achieved a Krippendorff's alpha agreement coefficient of 0.7081 in prospective testing with pharmacists. Some degree of automation could both accelerate the availability of allergy information and reduce the efforts for human coding.

Year: 2022

Citations: -

Full text: Yes

1. **Deep Learning to Classify Radiology Free-Text Reports**

Abstract: -

Year: 2017

Citations: 188

Full text: Yes

1. **Predicting Postoperative Mortality With Deep Neural Networks and Natural Language Processing: Model Development and Validation.**

Abstract: BACKGROUND: Machine learning (ML) achieves better predictions of postoperative mortality than previous prediction tools. Free-text descriptions of the preoperative diagnosis and the planned procedure are available preoperatively. Because reading these descriptions helps anesthesiologists evaluate the risk of the surgery, we hypothesized that deep learning (DL) models with unstructured text could improve postoperative mortality prediction. However, it is challenging to extract meaningful concept embeddings from this unstructured clinical text. OBJECTIVE: This study aims to develop a fusion DL model containing structured and unstructured features to predict the in-hospital 30-day postoperative mortality before surgery. ML models for predicting postoperative mortality using preoperative data with or without free clinical text were assessed. METHODS: We retrospectively collected preoperative anesthesia assessments, surgical information, and discharge summaries of patients undergoing general and neuraxial anesthesia from electronic health records (EHRs) from 2016 to 2020. We first compared the deep neural network (DNN) with other models using the same input features to demonstrate effectiveness. Then, we combined the DNN model with bidirectional encoder representations from transformers (BERT) to extract information from clinical texts. The effects of adding text information on the model performance were compared using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Statistical significance was evaluated using P<.05. RESULTS: The final cohort contained 121,313 patients who underwent surgeries. A total of 1562 (1.29%) patients died within 30 days of surgery. Our BERT-DNN model achieved the highest AUROC (0.964, 95% CI 0.961-0.967) and AUPRC (0.336, 95% CI 0.276-0.402). The AUROC of the BERT-DNN was significantly higher compared to logistic regression (AUROC=0.952, 95% CI 0.949-0.955) and the American Society of Anesthesiologist Physical Status (ASAPS AUROC=0.892, 95% CI 0.887-0.896) but not significantly higher compared to the DNN (AUROC=0.959, 95% CI 0.956-0.962) and the random forest (AUROC=0.961, 95% CI 0.958-0.964). The AUPRC of the BERT-DNN was significantly higher compared to the DNN (AUPRC=0.319, 95% CI 0.260-0.384), the random forest (AUPRC=0.296, 95% CI 0.239-0.360), logistic regression (AUPRC=0.276, 95% CI 0.220-0.339), and the ASAPS (AUPRC=0.149, 95% CI 0.107-0.203). CONCLUSIONS: Our BERT-DNN model has an AUPRC significantly higher compared to previously proposed models using no text and an AUROC significantly higher compared to logistic regression and the ASAPS. This technique helps identify patients with higher risk from the surgical description text in EHRs.

Year: 2022

Citations: -

Full text: Yes

1. **Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning.**

Abstract: BACKGROUND: The International Classification of Diseases (ICD) code is widely used as the reference in medical system and billing purposes. However, classifying diseases into ICD codes still mainly relies on humans reading a large amount of written material as the basis for coding. Coding is both laborious and time-consuming. Since the conversion of ICD-9 to ICD-10, the coding task became much more complicated, and deep learning- and natural language processing-related approaches have been studied to assist disease coders. OBJECTIVE: This paper aims at constructing a deep learning model for ICD-10 coding, where the model is meant to automatically determine the corresponding diagnosis and procedure codes based solely on free-text medical notes to improve accuracy and reduce human effort. METHODS: We used diagnosis records of the National Taiwan University Hospital as resources and apply natural language processing techniques, including global vectors, word to vectors, embeddings from language models, bidirectional encoder representations from transformers, and single head attention recurrent neural network, on the deep neural network architecture to implement ICD-10 auto-coding. Besides, we introduced the attention mechanism into the classification model to extract the keywords from diagnoses and visualize the coding reference for training freshmen in ICD-10. Sixty discharge notes were randomly selected to examine the change in the F(1)-score and the coding time by coders before and after using our model. RESULTS: In experiments on the medical data set of National Taiwan University Hospital, our prediction results revealed F(1)-scores of 0.715 and 0.618 for the ICD-10 Clinical Modification code and Procedure Coding System code, respectively, with a bidirectional encoder representations from transformers embedding approach in the Gated Recurrent Unit classification model. The well-trained models were applied on the ICD-10 web service for coding and training to ICD-10 users. With this service, coders can code with the F(1)-score significantly increased from a median of 0.832 to 0.922 (P<.05), but not in a reduced interval. CONCLUSIONS: The proposed model significantly improved the F(1)-score but did not decrease the time consumed in coding by disease coders.

Year: 2021

Citations: 11

Full text: Yes

1. **Multiple Sclerosis Severity Classification From Clinical Text**

Abstract: Multiple Sclerosis (MS) is a chronic, inflammatory and degenerative neurological disease, which is monitored by a specialist using the Expanded Disability Status Scale (EDSS) and recorded in unstructured text in the form of a neurology consult note. An EDSS measurement contains an overall "EDSS" score and several functional subscores. Typically, expert knowledge is required to interpret consult notes and generate these scores. Previous approaches used limited context length Word2Vec embeddings and keyword searches to predict scores given a consult note, but often failed when scores were not explicitly stated. In this work, we present MS-BERT, the first publicly available transformer model trained on real clinical data other than MIMIC. Next, we present MSBC, a classifier that applies MS-BERT to generate embeddings and predict EDSS and functional subscores. Lastly, we explore combining MSBC with other models through the use of Snorkel to generate scores for unlabelled consult notes. MSBC achieves state-of-the-art performance on all metrics and prediction tasks and outperforms the models generated from the Snorkel ensemble. We improve Macro-F1 by 0.12 (to 0.88) for predicting EDSS and on average by 0.29 (to 0.63) for predicting functional subscores over previous Word2Vec CNN and rule-based approaches.

Year: 2020

Citations: 5

Full text: Yes

1. **Transformer-based models for ICD-10 coding of death certificates with Portuguese text.**

Abstract: Natural Language Processing (NLP) can offer important tools for unlocking relevant information from clinical narratives. Although Transformer-based models can achieve remarkable results in several different NLP tasks, these models have been less used in clinical NLP, and particularly in low resource languages, of which Portuguese is one example. It is still not entirely clear whether pre-trained Transformer models are useful for clinical tasks, without further architecture engineering or particular training strategies. In this work, we propose a BERT model to assign ICD-10 codes for causes of death, by analyzing free-text descriptions in death certificates, together with the associated autopsy reports and clinical bulletins, from the Portuguese Ministry of Health. We used a novel pre-training procedure that incorporates in-domain knowledge, and also a fine-tuning method to address the class imbalance issue. Experimental results show that, in this particular clinical task that requires the processing of relatively short documents, Transformer-based models can achieve very strong results, significantly outperforming tailored approaches based on recurrent neural networks.

Year: 2022

Citations: -

Full text: Yes

1. **Multi-label Classification for Clinical Text with Feature-level Attention**

Abstract: Multi-label text classification, which tags a given plain text with the most relevant labels from a label space, is an important task in the natural language process. To diagnose diseases, clinical researchers use a machine-learning algorithm to do multi-label clinical text classification. However, conventional machine learning methods can neither capture deep semantic information nor the context of words strictly. Diagnostic information from the EHRs (Electronic Health Records) is mainly constructed by unstructured clinical free text which is an obstacle for clinical feature extraction. Moreover, feature engineering is time-consuming and labor-intensive. With the rapid development of deep learning, we apply neural network models to resolve this problem mentioned above. To favor multi-label classification on EHRs, we propose FAMLC-BERT (Feature-level Attention for Multi-label classification on BERT) to capture semantic features from different layers. The model uses feature-level attention with BERT to recognize the labels of EHRs. We empirically compared our model with other state-of-the-art models on real-world documents collected from the hospital. Experiments show that our model achieved significant improvements compared to other selected benchmarks.

Year: 2020

Citations: 4

Full text: Yes

1. **BERTje: A Dutch BERT Model**

Abstract: The transformer-based pre-trained language model BERT has helped to improve state-of-the-art performance on many natural language processing (NLP) tasks. Using the same architecture and parameters, we developed and evaluated a monolingual Dutch BERT model called BERTje. Compared to the multilingual BERT model, which includes Dutch but is only based on Wikipedia text, BERTje is based on a large and diverse dataset of 2.4 billion tokens. BERTje consistently outperforms the equally-sized multilingual BERT model on downstream NLP tasks (part-of-speech tagging, named-entity recognition, semantic role labeling, and sentiment analysis). Our pre-trained Dutch BERT model is made available at <https://github.com/wietsedv/bertje>.

Year: 2019

Citations: 215

Full text: Yes

1. **Optimal Subarchitecture Extraction For BERT**

Abstract: We extract an optimal subset of architectural parameters for the BERT architecture from Devlin et al. (2018) by applying recent breakthroughs in algorithms for neural architecture search. This optimal subset, which we refer to as "Bort", is demonstrably smaller, having an effective (that is, not counting the embedding layer) size of $5.5\%$ the original BERT-large architecture, and $16\%$ of the net size. Bort is also able to be pretrained in $288$ GPU hours, which is $1.2\%$ of the time required to pretrain the highest-performing BERT parametric architectural variant, RoBERTa-large (Liu et al., 2019), and about $33\%$ of that of the world-record, in GPU hours, required to train BERT-large on the same hardware. It is also $7.9$x faster on a CPU, as well as being better performing than other compressed variants of the architecture, and some of the non-compressed variants: it obtains performance improvements of between $0.3\%$ and $31\%$, absolute, with respect to BERT-large, on multiple public natural language understanding (NLU) benchmarks.

Year: 2020

Citations: 17

Full text: Yes

1. **RobBERT: a Dutch RoBERTa-based Language Model**

Abstract: Pre-trained language models have been dominating the field of natural language processing in recent years, and have led to significant performance gains for various complex natural language tasks. One of the most prominent pre-trained language models is BERT, which was released as an English as well as a multilingual version. Although multilingual BERT performs well on many tasks, recent studies show that BERT models trained on a single language significantly outperform the multilingual version. Training a Dutch BERT model thus has a lot of potential for a wide range of Dutch NLP tasks. While previous approaches have used earlier implementations of BERT to train a Dutch version of BERT, we used RoBERTa, a robustly optimized BERT approach, to train a Dutch language model called RobBERT. We measured its performance on various tasks as well as the importance of the fine-tuning dataset size. We also evaluated the importance of language-specific tokenizers and the model's fairness. We found that RobBERT improves state-of-the-art results for various tasks, and especially significantly outperforms other models when dealing with smaller datasets. These results indicate that it is a powerful pre-trained model for a large variety of Dutch language tasks. The pre-trained and fine-tuned models are publicly available to support further downstream Dutch NLP applications.

Year: 2020

Citations: 141

Full text: Yes

1. **RobBERTje: a Distilled Dutch BERT Model**

Abstract: Pre-trained large-scale language models such as BERT have gained a lot of attention thanks to their outstanding performance on a wide range of natural language tasks. However, due to their large number of parameters, they are resource-intensive both to deploy and to fine-tune. Researchers have created several methods for distilling language models into smaller ones to increase efficiency, with a small performance trade-off. In this paper, we create several different distilled versions of the state-of-the-art Dutch RobBERT model and call them RobBERTje. The distillations differ in their distillation corpus, namely whether or not they are shuffled and whether they are merged with subsequent sentences. We found that the performance of the models using the shuffled versus non-shuffled datasets is similar for most tasks and that randomly merging subsequent sentences in a corpus creates models that train faster and perform better on tasks with long sequences. Upon comparing distillation architectures, we found that the larger DistilBERT architecture worked significantly better than the Bort hyperparametrization. Interestingly, we also found that the distilled models exhibit less gender-stereotypical bias than its teacher model. Since smaller architectures decrease the time to fine-tune, these models allow for more efficient training and more lightweight deployment of many Dutch downstream language tasks.

Year: 2022

Citations: 5

Full text: Yes

1. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Abstract: We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

Year: 2018

Citations: 59.487

Full text: Yes

1. **Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text.**

Abstract: INTRODUCTION: Proximal humeral fractures account for a significant proportion of all fractures. Detailed accurate classification of the type and severity of the fracture is a key component of clinical decision making, treatment and plays an important role in orthopaedic trauma research. This research aimed to assess the performance of Machine Learning (ML) multiclass classification algorithms to classify proximal humeral fractures using radiology text data. MATERIALS AND METHODS: Data from adult (16 + years) patients admitted to a major trauma centre for management of their proximal humerus fracture from January 2010 to January 2019 were used (1,324). Six input text datasets were used for classification: X-ray and/or CT scan reports (primary) and concatenation of patient age and/or patient sex. One of seven Neer class labels were classified. Models were evaluated using accuracy, recall, precision, F1, and One-versus-rest scores. RESULTS: A number of statistical ML algorithms performed acceptably and one of the BERT models, exhibiting good accuracy of 61% and an excellent one-versus-rest score above 0.8. The highest precision, recall and F1 scores were 50%, 39% and 39% respectively, being considered reasonable scores with the sparse text data used and in the context of machine learning. CONCLUSION: ML and BERT algorithms based on routine unstructured X-ray and CT text reports, combined with the demographics of the patient, show promise in Neer classification of proximal humeral fractures to aid research. Use of these algorithms shows potential to speed up the classification task and assist radiologist, surgeons and researchers.

Year: 2022

Citations: -

Full text: Yes

1. **Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records.**

Abstract: OBJECTIVE: Pituitary adenomas are the most common type of pituitary disorders, which usually occur in young adults and often affect the patient's physical development, labor capacity and fertility. Clinical free texts noted in electronic medical records (EMRs) of pituitary adenomas patients contain abundant diagnosis and treatment information. However, this information has not been well utilized because of the challenge to extract information from unstructured clinical texts. This study aims to enable machines to intelligently process clinical information, and automatically extract clinical named entity for pituitary adenomas from Chinese EMRs. METHODS: The clinical corpus used in this study was from one pituitary adenomas neurosurgery treatment center of a 3A hospital in China. Four types of fine-grained texts of clinical records were selected, which included notes from present illness, past medical history, case characteristics and family history of 500 pituitary adenoma inpatients. The dictionary-based matching, conditional random fields (CRF), bidirectional long short-term memory with CRF (BiLSTM-CRF), and bidirectional encoder representations from transformers with BiLSTM-CRF (BERT-BiLSTM-CRF) were used to extract clinical entities from a Chinese EMRs corpus. A comprehensive dictionary was constructed based on open source vocabularies and a domain dictionary for pituitary adenomas to conduct the dictionary-based matching method. We selected features such as part of speech, radical, document type, and the position of characters to train the CRF-based model. Random character embeddings and the character embeddings pretrained by BERT were used respectively as the input features for the BiLSTM-CRF model and the BERT-BiLSTM-CRF model. Both strict metric and relaxed metric were used to evaluate the performance of these methods. RESULTS: Experimental results demonstrated that the deep learning and other machine learning methods were able to automatically extract clinical named entities, including symptoms, body regions, diseases, family histories, surgeries, medications, and disease courses of pituitary adenomas from Chinese EMRs. With regard to overall performance, BERT-BiLSTM-CRF has the highest strict F1 value of 91.27% and the highest relaxed F1 value of 95.57% respectively. Additional evaluations showed that BERT-BiLSTM-CRF performed best in almost all entity recognition except surgery and disease course. BiLSTM-CRF performed best in disease course entity recognition, and performed as well as the CRF model for part of speech, radical and document type features, with both strict and relaxed F1 value reaching 96.48%. The CRF model with part of speech, radical and document type features performed best in surgery entity recognition with relaxed F1 value of 95.29%. CONCLUSIONS: In this study, we conducted four entity recognition methods for pituitary adenomas based on Chinese EMRs. It demonstrates that the deep learning methods can effectively extract various types of clinical entities with satisfying performance. This study contributed to the clinical named entity extraction from Chinese neurosurgical EMRs. The findings could also assist in information extraction in other Chinese medical texts.

Year: 2022

Citations: 2

Full text: Yes

1. **Natural Language Processing for Automated Classification of Qualitative Data From Interviews of Patients With Cancer.**

Abstract: OBJECTIVES: This study sought to explore the use of novel natural language processing (NLP) methods for classifying unstructured, qualitative textual data from interviews of patients with cancer to identify patient-reported symptoms and impacts on quality of life. METHODS: We tested the ability of 4 NLP models to accurately classify text from interview transcripts as "symptom," "quality of life impact," and "other." Interview data sets from patients with hepatocellular carcinoma (HCC) (n = 25), biliary tract cancer (BTC) (n = 23), and gastric cancer (n = 24) were used. Models were cross-validated with transcript subsets designated for training, validation, and testing. Multiclass classification performance of the 4 models was evaluated at paragraph and sentence level using the HCC testing data set and analyzed by the one-versus-rest technique quantified by the receiver operating characteristic area under the curve (ROC AUC) score. RESULTS: NLP models accurately classified multiclass text from patient interviews. The Bidirectional Encoder Representations from Transformers model generally outperformed all other models at paragraph and sentence level. The highest predictive performance of the Bidirectional Encoder Representations from Transformers model was observed using the HCC data set to train and BTC data set to test (mean ROC AUC, 0.940 [SD 0.028]), with similarly high predictive performance using balanced and imbalanced training data sets from BTC and gastric cancer populations. CONCLUSIONS: NLP models were accurate in predicting multiclass classification of text from interviews of patients with cancer, with most surpassing 0.9 ROC AUC at paragraph level. NLP may be a useful tool for scaling up processing of patient interviews in clinical studies and, thus, could serve to facilitate patient input into drug development and improving patient care.

Year: 2022

Citations: -

Full text: Yes

1. **Deep Learning-based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports.**

Abstract: PURPOSE: To train a deep natural language processing (NLP) model, using data mined structured oncology reports (SOR), for rapid tumor response category (TRC) classification from free-text oncology reports (FTOR) and to compare its performance with human readers and conventional NLP algorithms. MATERIALS AND METHODS: In this retrospective study, databases of three independent radiology departments were queried for SOR and FTOR dated from March 2018 to August 2021. An automated data mining and curation pipeline was developed to extract Response Evaluation Criteria in Solid Tumors-related TRCs for SOR for ground truth definition. The deep NLP bidirectional encoder representations from transformers (BERT) model and three feature-rich algorithms were trained on SOR to predict TRCs in FTOR. Models' F1 scores were compared against scores of radiologists, medical students, and radiology technologist students. Lexical and semantic analyses were conducted to investigate human and model performance on FTOR. RESULTS: Oncologic findings and TRCs were accurately mined from 9653 of 12 833 (75.2%) queried SOR, yielding oncology reports from 10 455 patients (mean age, 60 years ± 14 [SD]; 5303 women) who met inclusion criteria. On 802 FTOR in the test set, BERT achieved better TRC classification results (F1, 0.70; 95% CI: 0.68, 0.73) than the best-performing reference linear support vector classifier (F1, 0.63; 95% CI: 0.61, 0.66) and technologist students (F1, 0.65; 95% CI: 0.63, 0.67), had similar performance to medical students (F1, 0.73; 95% CI: 0.72, 0.75), but was inferior to radiologists (F1, 0.79; 95% CI: 0.78, 0.81). Lexical complexity and semantic ambiguities in FTOR influenced human and model performance, revealing maximum F1 score drops of -0.17 and -0.19, respectively. CONCLUSION: The developed deep NLP model reached the performance level of medical students but not radiologists in curating oncologic outcomes from radiology FTOR.Keywords: Neural Networks, Computer Applications-Detection/Diagnosis, Oncology, Research Design, Staging, Tumor Response, Comparative Studies, Decision Analysis, Experimental Investigations, Observer Performance, Outcomes Analysis Supplemental material is available for this article. © RSNA, 2022.

Year: 2022

Citations: 4

Full text: Yes

1. **A hybrid model to identify fall occurrence from electronic health records.**

Abstract: INTRODUCTION: Falls are a leading cause of unintentional injury in the elderly. Electronic health records (EHRs) offer the unique opportunity to develop models that can identify fall events. However, identifying fall events in clinical notes requires advanced natural language processing (NLP) to simultaneously address multiple issues because the word "fall" is a typical homonym. METHODS: We implemented a context-aware language model, Bidirectional Encoder Representations from Transformers (BERT) to identify falls from the EHR text and further fused the BERT model into a hybrid architecture coupled with post-hoc heuristic rules to enhance the performance. The models were evaluated on real world EHR data and were compared to conventional rule-based and deep learning models (CNN and Bi-LSTM). To better understand the ability of each approach to identify falls, we further categorize fall-related concepts (i.e., risk of fall, prevention of fall, homonym) and performed a detailed error analysis. RESULTS: The hybrid model achieved the highest f1-score on sentence (0.971), document (0.985), and patient (0.954) level. At the sentence level (basic data unit in the model), the hybrid model had 0.954, 1.000, 0.988, and 0.999 in sensitivity, specificity, positive predictive value, and negative predictive value, respectively. The error analysis showed that that machine learning-based approaches demonstrated higher performance than a rule-based approach in challenging cases that required contextual understanding. The context-aware language model (BERT) slightly outperformed the word embedding approach trained on Bi-LSTM. No single model yielded the best performance for all fall-related semantic categories. CONCLUSION: A context-aware language model (BERT) was able to identify challenging fall events that requires context understanding in EHR free text. The hybrid model combined with post-hoc rules allowed a custom fix on the BERT outcomes and further improved the performance of fall detection.

Year: 2022

Citations: 1

Full text: Yes

1. **DeepNote-GNN: Predicting Hospital Readmission Using Clinical Notes and Patient Network**

Abstract: With the increasing availability of Electronic Health Records (EHRs) and advances in deep learning techniques, developing deep predictive models that use EHR data to solve healthcare problems has gained momentum in recent years. The majority of clinical predictive models benefit from structured data in EHR (e.g., lab measurements and medications). Still, learning clinical outcomes from all possible information sources is one of the main challenges when building predictive models. This work focuses on two sources of information that have been underused by researchers; unstructured data (e.g., clinical notes) and a patient network. We propose a novel hybrid deep learning model, DeepNote-GNN, that integrates clinical notes information and patient network topological structure to improve 30-day hospital readmission prediction. DeepNote-GNN is a robust deep learning framework consisting of two modules: DeepNote and patient network. DeepNote extracts deep representations of clinical notes using a feature aggregation unit on top of a state-of-the-art Natural Language Processing (NLP) technique - BERT. By exploiting these deep representations, a patient network is built, and Graph Neural Network (GNN) is used to train the network for hospital readmission predictions. Performance evaluation on the MIMIC-III dataset demonstrates that DeepNote-GNN achieves superior results compared to the state-of-the-art baselines on the 30-day hospital readmission task. We extensively analyze the DeepNote-GNN model to illustrate the effectiveness and contribution of each component of it. The model analysis shows that patient network has a significant contribution to the overall performance, and DeepNote-GNN is robust and can consistently perform well on the 30-day readmission prediction task.

Year: 2021

Citations: 9

Full text: Yes

1. **Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing**

Abstract: Pretraining large neural language models, such as BERT, has led to impressive gains on many natural language processing (NLP) tasks. However, most pretraining efforts focus on general domain corpora, such as newswire and Web. A prevailing assumption is that even domain-specific pretraining can benefit by starting from general-domain language models. In this article, we challenge this assumption by showing that for domains with abundant unlabeled text, such as biomedicine, pretraining language models from scratch results in substantial gains over continual pretraining of general-domain language models. To facilitate this investigation, we compile a comprehensive biomedical NLP benchmark from publicly available datasets. Our experiments show that domain-specific pretraining serves as a solid foundation for a wide range of biomedical NLP tasks, leading to new state-of-the-art results across the board. Further, in conducting a thorough evaluation of modeling choices, both for pretraining and task-specific fine-tuning, we discover that some common practices are unnecessary with BERT models, such as using complex tagging schemes in named entity recognition. To help accelerate research in biomedical NLP, we have released our state-of-the-art pretrained and task-specific models for the community, and created a leaderboard featuring our BLURB benchmark (short for Biomedical Language Understanding & Reasoning Benchmark) at https://aka.ms/BLURB .

Year: 2022

Citations: 598

Full text: Yes

1. **Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing.**

Abstract: OBJECTIVE: Social determinants of health (SDOH) are non-medical factors that can profoundly impact patient health outcomes. However, SDOH are rarely available in structured electronic health record (EHR) data such as diagnosis codes, and more commonly found in unstructured narrative clinical notes. Hence, identifying social context from unstructured EHR data has become increasingly important. Yet, previous work on using natural language processing to automate extraction of SDOH from text (a) usually focuses on an ad hoc selection of SDOH, and (b) does not use the latest advances in deep learning. Our objective was to advance automatic extraction of SDOH from clinical text by (a) systematically creating a set of SDOH based on standard biomedical and psychiatric ontologies, and (b) training state-of-the-art deep neural networks to extract mentions of these SDOH from clinical notes. DESIGN: A retrospective cohort study. SETTING AND PARTICIPANTS: Data were extracted from the Medical Information Mart for Intensive Care (MIMIC-III) database. The corpus comprised 3,504 social related sentences from 2,670 clinical notes. METHODS: We developed a framework for automated classification of multiple SDOH categories. Our dataset comprised narrative clinical notes under the "Social Work" category in the MIMIC-III Clinical Database. Using standard terminologies, SNOMED-CT and DSM-IV, we systematically curated a set of 13 SDOH categories and created annotation guidelines for these. After manually annotating the 3,504 sentences, we developed and tested three deep neural network (DNN) architectures - convolutional neural network (CNN), long short-term memory (LSTM) network, and the Bidirectional Encoder Representations from Transformers (BERT) - for automated detection of eight SDOH categories. We also compared these DNNs to three baselines models: (1) cTAKES, as well as (2) L2-regularized logistic regression and (3) random forests on bags-of-words. Model evaluation metrics included micro- and macro- F1, and area under the receiver operating characteristic curve (AUC). RESULTS: All three DNN models accurately classified all SDOH categories (minimum micro-F1 = 0.632, minimum macro-AUC = 0.854). Compared to the CNN and LSTM, BERT performed best in most key metrics (micro-F1 = 0.690, macro-AUC = 0.907). The BERT model most effectively identified the "occupational" category (F1 = 0.774, AUC = 0.965) and least effectively identified the "non-SDOH" category (F = 0.491, AUC = 0.788). BERT outperformed cTAKES in distinguishing social vs non-social sentences (BERT F1 = 0.87 vs. cTAKES F1 = 0.06), and outperformed logistic regression (micro-F1 = 0.649, macro-AUC = 0.696) and random forest (micro-F1 = 0.502, macro-AUC = 0.523) trained on bag-of-words. CONCLUSIONS: Our study framework with DNN models demonstrated improved performance for efficiently identifying a systematic range of SDOH categories from clinical notes in the EHR. Improved identification of patient SDOH may further improve healthcare outcomes.

Year: 2022

Citations: 12

Full text: Yes

1. **Prospects for Dutch Emotion Detection: Insights from the New EmotioNL Dataset**

Abstract: Although emotion detection has become a crucial research direction in NLP, the main focus is on English resources and data. The main obstacles for more specialized emotion detection are the lack of annotated data in smaller languages and the limited emotion taxonomy. In a first step towards improving emotion detection for Dutch, we present EmotioNL, an emotion dataset consisting of 1,000 Dutch tweets and 1,000 captions from TV-shows, annotated with emotion categories (anger, fear, joy, love, sadness and neutral) and dimensions (valence, arousal and dominance). We evaluate the state-of-the-art Dutch transformer models BERTje and RobBERT on this new dataset, investigate model generalizability across domains and perform a thorough error analysis based on the Component Process Model of emotions.

Year: 2022

Citations: 3

Full text: Yes

1. **Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach.**

Abstract: BACKGROUND: Lung cancer is the leading cause of cancer deaths worldwide. Clinical staging of lung cancer plays a crucial role in making treatment decisions and evaluating prognosis. However, in clinical practice, approximately one-half of the clinical stages of lung cancer patients are inconsistent with their pathological stages. As one of the most important diagnostic modalities for staging, chest computed tomography (CT) provides a wealth of information about cancer staging, but the free-text nature of the CT reports obstructs their computerization. OBJECTIVE: We aimed to automatically extract the staging-related information from CT reports to support accurate clinical staging of lung cancer. METHODS: In this study, we developed an information extraction (IE) system to extract the staging-related information from CT reports. The system consisted of the following three parts: named entity recognition (NER), relation classification (RC), and postprocessing (PP). We first summarized 22 questions about lung cancer staging based on the TNM staging guideline. Next, three state-of-the-art NER algorithms were implemented to recognize the entities of interest. Next, we designed a novel RC method using the relation sign constraint (RSC) to classify the relations between entities. Finally, a rule-based PP module was established to obtain the formatted answers using the results of NER and RC. RESULTS: We evaluated the developed IE system on a clinical data set containing 392 chest CT reports collected from the Department of Thoracic Surgery II in the Peking University Cancer Hospital. The experimental results showed that the bidirectional encoder representation from transformers (BERT) model outperformed the iterated dilated convolutional neural networks-conditional random field (ID-CNN-CRF) and bidirectional long short-term memory networks-conditional random field (Bi-LSTM-CRF) for NER tasks with macro-F1 scores of 80.97% and 90.06% under the exact and inexact matching schemes, respectively. For the RC task, the proposed RSC showed better performance than the baseline methods. Further, the BERT-RSC model achieved the best performance with a macro-F1 score of 97.13% and a micro-F1 score of 98.37%. Moreover, the rule-based PP module could correctly obtain the formatted results using the extractions of NER and RC, achieving a macro-F1 score of 94.57% and a micro-F1 score of 96.74% for all the 22 questions. CONCLUSIONS: We conclude that the developed IE system can effectively and accurately extract information about lung cancer staging from CT reports. Experimental results show that the extracted results have significant potential for further use in stage verification and prediction to facilitate accurate clinical staging.

Year: 2021

Citations: 5

Full text: Yes

1. **Predicting Glaucoma Progression Requiring Surgery Using Clinical Free-Text Notes and Transfer Learning With Transformers.**

Abstract: PURPOSE: We evaluated the use of massive transformer-based language models to predict glaucoma progression requiring surgery using ophthalmology clinical notes from electronic health records (EHRs). METHODS: Ophthalmology clinical notes for 4512 glaucoma patients at a single center from 2008 to 2020 were identified from the EHRs. Four different pre-trained Bidirectional Encoder Representations from Transformers (BERT)-based models were fine-tuned on ophthalmology clinical notes from the patients' first 120 days of follow-up for the task of predicting which patients would require glaucoma surgery. Models were evaluated with standard metrics, including area under the receiver operating characteristic curve (AUROC) and F1 score. RESULTS: Of the patients, 748 progressed to require glaucoma surgery (16.6%). The original BERT model had the highest AUROC (73.4%; F1 = 45.0%) for identifying these patients, followed by RoBERTa, with an AUROC of 72.4% (F1 = 44.7%); DistilBERT, with an AUROC of 70.2% (F1 = 42.5%); and BioBERT, with an AUROC of 70.1% (F1 = 41.7%). All models had higher F1 scores than an ophthalmologist's review of clinical notes (F1 = 29.9%). CONCLUSIONS: Using transfer learning with massively pre-trained BERT-based models is a natural language processing approach that can access the wealth of clinical information stored within ophthalmology clinical notes to predict the progression of glaucoma. Future work to improve model performance can focus on integrating structured or imaging data or further tailoring the BERT models to ophthalmology domain-specific text. TRANSLATIONAL RELEVANCE: Predictive models can provide the basis for clinical decision support tools to aid clinicians in identifying high- or low-risk patients to maximally tailor glaucoma treatments.

Year: 2022

Citations: 2

Full text: Yes

1. **ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission**

Abstract: Clinical notes contain information about patients that goes beyond structured data like lab values and medications. However, clinical notes have been underused relative to structured data, because notes are high-dimensional and sparse. This work develops and evaluates representations of clinical notes using bidirectional transformers (ClinicalBERT). ClinicalBERT uncovers high-quality relationships between medical concepts as judged by humans. ClinicalBert outperforms baselines on 30-day hospital readmission prediction using both discharge summaries and the first few days of notes in the intensive care unit. Code and model parameters are available.

Year: 2020

Citations: 430

Full text: Yes

1. **Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training**

Abstract: Recently a number of studies demonstrated impressive performance on diverse vision-language multi-modal tasks such as image captioning and visual question answering by extending the BERT architecture with multi-modal pre-training objectives. In this work we explore a broad set of multi-modal representation learning tasks in the medical domain, specifically using radiology images and the unstructured report. We propose Medical Vision Language Learner (MedViLL), which adopts a BERT-based architecture combined with a novel multi-modal attention masking scheme to maximize generalization performance for both vision-language understanding tasks (diagnosis classification, medical image-report retrieval, medical visual question answering) and vision-language generation task (radiology report generation). By statistically and rigorously evaluating the proposed model on four downstream tasks with three radiographic image-report datasets (MIMIC-CXR, Open-I, and VQA-RAD), we empirically demonstrate the superior downstream task performance of MedViLL against various baselines, including task-specific architectures.

Year: 2022

Citations: 16

Full text: Yes

1. **VisualCheXbert: Addressing the Discrepancy between Radiology Report Labels and Image Labels**

Abstract: Automatic extraction of medical conditions from free-text radiology reports is critical for supervising computer vision models to interpret medical images. In this work, we show that radiologists labeling reports significantly disagree with radiologists labeling corresponding chest X-ray images, which reduces the quality of report labels as proxies for image labels. We develop and evaluate methods to produce labels from radiology reports that have better agreement with radiologists labeling images. Our best performing method, called VisualCheXbert, uses a biomedically-pretrained BERT model to directly map from a radiology report to the image labels, with a supervisory signal determined by a computer vision model trained to detect medical conditions from chest X-ray images. We find that VisualCheXbert outperforms an approach using an existing radiology report labeler by an average F1 score of 0.14 (95% CI 0.12, 0.17). We also find that VisualCheXbert better agrees with radiologists labeling chest X-ray images than do radiologists labeling the corresponding radiology reports by an average F1 score across several medical conditions of between 0.12 (95% CI 0.09, 0.15) and 0.21 (95% CI 0.18, 0.24).

Year: 2021

Citations: 13

Full text: Yes

1. **RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification.**

Abstract: Radiology reports are unstructured and contain the imaging findings and corresponding diagnoses transcribed by radiologists which include clinical facts and negated and/or uncertain statements. Extracting pathologic findings and diagnoses from radiology reports is important for quality control, population health, and monitoring of disease progress. Existing works, primarily rely either on rule-based systems or transformer-based pre-trained model fine-tuning, but could not take the factual and uncertain information into consideration, and therefore generate false positive outputs. In this work, we introduce three sedulous augmentation techniques which retain factual and critical information while generating augmentations for contrastive learning. We introduce RadBERT-CL, which fuses these information into BlueBert via a self-supervised contrastive loss. Our experiments on MIMIC-CXR show superior performance of RadBERT-CL on fine-tuning for multi-class, multi-label report classification. We illustrate that when few labeled data are available, RadBERT-CL outperforms conventional SOTA transformers (BERT/BlueBert) by significantly larger margins (6-11%). We also show that the representations learned by RadBERT-CL can capture critical medical information in the latent space.

Year: 2021

Citations: 4

Full text: Yes

1. **TinyBERT: Distilling BERT for Natural Language Understanding**

Abstract: Language model pre-training, such as BERT, has significantly improved the performances of many natural language processing tasks. However, pre-trained language models are usually computationally expensive, so it is difficult to efficiently execute them on resource-restricted devices. To accelerate inference and reduce model size while maintaining accuracy, we first propose a novel Transformer distillation method that is specially designed for knowledge distillation (KD) of the Transformer-based models. By leveraging this new KD method, the plenty of knowledge encoded in a large teacher BERT can be effectively transferred to a small student Tiny-BERT. Then, we introduce a new two-stage learning framework for TinyBERT, which performs Transformer distillation at both the pretraining and task-specific learning stages. This framework ensures that TinyBERT can capture he general-domain as well as the task-specific knowledge in BERT.

TinyBERT with 4 layers is empirically effective and achieves more than 96.8% the performance of its teacher BERTBASE on GLUE benchmark, while being 7.5x smaller and 9.4x faster on inference. TinyBERT with 4 layers is also significantly better than 4-layer state-of-the-art baselines on BERT distillation, with only about 28% parameters and about 31% inference time of them. Moreover, TinyBERT with 6 layers performs on-par with its teacher BERTBASE.

Year: 2019

Citations: 951

Full text: Yes

1. **SpanBERT: Improving Pre-training by Representing and Predicting Spans**

Abstract: We present SpanBERT, a pre-training method that is designed to better represent and predict spans of text. Our approach extends BERT by (1) masking contiguous random spans, rather than random tokens, and (2) training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it. SpanBERT consistently outperforms BERT and our better-tuned baselines, with substantial gains on span selection tasks such as question answering and coreference resolution. In particular, with the same training data and model size as BERT-large, our single model obtains 94.6% and 88.7% F1 on SQuAD 1.1 and 2.0, respectively. We also achieve a new state of the art on the OntoNotes coreference resolution task (79.6\% F1), strong performance on the TACRED relation extraction benchmark, and even show gains on GLUE.

Year: 2019

Citations: 1.306

Full text: Yes

1. **Use of artificial intelligence to identify data elements for The Japanese Orthopaedic Association National Registry from operative records.**

Abstract: BACKGROUND: The Japanese Orthopaedic Association National Registry (JOANR) was recently launched in Japan and is expected to improve the quality of medical care. However, surgeons must register ten detailed features for total hip arthroplasty, which is labor intensive. One possible solution is to use a system that automatically extracts information about the surgeries. Although it is not easy to extract features from an operative record consisting of free-text data, natural language processing has been used to extract features from operative records. This study aimed to evaluate the best natural language processing method for building a system that automatically detects some elements in the JOANR from the operative records of total hip arthroplasty. METHODS: We obtained operative records of total hip arthroplasty (n = 2574) in three hospitals and targeted two items: surgical approach and fixation technique. We compared the accuracy of three natural language processing methods: rule-based algorithms, machine learning, and bidirectional encoder representations from transformers (BERT). RESULTS: In the surgical approach task, the accuracy of BERT was superior to that of the rule-based algorithm (99.6% vs. 93.6%, p < 0.001), comparable to machine learning. In the fixation technique task, the accuracy of BERT was superior to the rule-based algorithm and machine learning (96% vs. 74%, p < 0.0001 and 94%, p = 0.0004). CONCLUSIONS: BERT is the most appropriate method for building a system that automatically detects the surgical approach and fixation technique.

Year: 2022

Citations: -

Full text: No

1. **HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language**

Abstract: The detection of abusive or offensive remarks in social texts has received significant attention in research. In several related shared tasks, BERT has been shown to be the state-of-the-art. In this paper, we propose to utilize lexical features derived from a hate lexicon towards improving the performance of BERT in such tasks. We explore different ways to utilize the lexical features in the form of lexicon-based encodings at the sentence level or embeddings at the word level. We provide an extensive dataset evaluation that addresses in-domain as well as cross-domain detection of abusive content to render a complete picture. Our results indicate that our proposed models combining BERT with lexical features help improve over a baseline BERT model in many of our in-domain and cross-domain experiments.

Year: 2020

Citations: 33

Full text: Yes

1. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**

Abstract: Increasing model size when pretraining natural language representations often results in improved performance on downstream tasks. However, at some point further model increases become harder due to GPU/TPU memory limitations and longer training times. To address these problems, we present two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT. Comprehensive empirical evidence shows that our proposed methods lead to models that scale much better compared to the original BERT. We also use a self-supervised loss that focuses on modeling inter-sentence coherence, and show it consistently helps downstream tasks with multi-sentence inputs. As a result, our best model establishes new state-of-the-art results on the GLUE, RACE, and \squad benchmarks while having fewer parameters compared to BERT-large. The code and the pretrained models are available at <https://github.com/google-research/ALBERT>.

Year: 2019

Citations: 4.285

Full text: Yes

1. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**

Abstract: MOTIVATION: Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from biomedical literature has gained popularity among researchers, and deep learning has boosted the development of effective biomedical text mining models. However, directly applying the advancements in NLP to biomedical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora. In this article, we investigate how the recently introduced pre-trained language model BERT can be adapted for biomedical corpora. RESULTS: We introduce BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. With almost the same architecture across tasks, BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. While BERT obtains performance comparable to that of previous state-of-the-art models, BioBERT significantly outperforms them on the following three representative biomedical text mining tasks: biomedical named entity recognition (0.62% F1 score improvement), biomedical relation extraction (2.80% F1 score improvement) and biomedical question answering (12.24% MRR improvement). Our analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts. AVAILABILITY AND IMPLEMENTATION: We make the pre-trained weights of BioBERT freely available at https://github.com/naver/biobert-pretrained, and the source code for fine-tuning BioBERT available at <https://github.com/dmis-lab/biobert>.

Year: 2020

Citations: 3.205

Full text: Yes

1. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**

Abstract: We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of the original sentences and using a novel in-filling scheme, where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD, achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pretraining. We also report ablation experiments that replicate other pretraining schemes within the BART framework, to better measure which factors most influence end-task performance.

Year: 2019

Citations: 4.410

Full text: Yes

1. **Does BERT need domain adaptation for clinical negation detection?**

Abstract: INTRODUCTION: Classifying whether concepts in an unstructured clinical text are negated is an important unsolved task. New domain adaptation and transfer learning methods can potentially address this issue. OBJECTIVE: We examine neural unsupervised domain adaptation methods, introducing a novel combination of domain adaptation with transformer-based transfer learning methods to improve negation detection. We also want to better understand the interaction between the widely used bidirectional encoder representations from transformers (BERT) system and domain adaptation methods. MATERIALS AND METHODS: We use 4 clinical text datasets that are annotated with negation status. We evaluate a neural unsupervised domain adaptation algorithm and BERT, a transformer-based model that is pretrained on massive general text datasets. We develop an extension to BERT that uses domain adversarial training, a neural domain adaptation method that adds an objective to the negation task, that the classifier should not be able to distinguish between instances from 2 different domains. RESULTS: The domain adaptation methods we describe show positive results, but, on average, the best performance is obtained by plain BERT (without the extension). We provide evidence that the gains from BERT are likely not additive with the gains from domain adaptation. DISCUSSION: Our results suggest that, at least for the task of clinical negation detection, BERT subsumes domain adaptation, implying that BERT is already learning very general representations of negation phenomena such that fine-tuning even on a specific corpus does not lead to much overfitting. CONCLUSION: Despite being trained on nonclinical text, the large training sets of models like BERT lead to large gains in performance for the clinical negation detection task.

Year: 2020

Citations: 35

Full text: Yes

1. **EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain**

Abstract: Transformer-based neural language models have led to breakthroughs for a variety of natural language processing (NLP) tasks. However, most models are pretrained on general domain data. We propose a methodology to produce a model focused on the clinical domain: continued pretraining of a model with a broad representation of biomedical terminology (PubMedBERT) on a clinical corpus along with a novel entity-centric masking strategy to infuse domain knowledge in the learning process. We show that such a model achieves superior results on clinical extraction tasks by comparing our entity-centric masking strategy with classic random masking on three clinical NLP tasks: cross-domain negation detection (Wu et al., 2014), document time relation (DocTimeRel) classification (Lin et al., 2020b), and temporal relation extraction (Wright-Bettner et al., 2020). We also evaluate our models on the PubMedQA(Jin et al., 2019) dataset to measure the models’ performance on a nonentity-centric task in the biomedical domain. The language addressed in this work is English. © 2021 Association for Computational Linguistics

Year: 2021

Citations: 13

Full text: Yes

1. **A disease-specific language representation model for cerebrovascular disease research.**

Abstract: BACKGROUND: Effectively utilizing disease-relevant text information from unstructured clinical notes for medical research presents many challenges. BERT (Bidirectional Encoder Representation from Transformers) related models such as BioBERT and ClinicalBERT, pre-trained on biomedical corpora and general clinical information, have shown promising performance in various biomedical language processing tasks. OBJECTIVES: This study aims to explore whether a BERT-based model pre-trained on disease-related clinical information can be more effective for cerebrovascular disease-relevant research. METHODS: This study proposed the StrokeBERT which was initialized from BioBERT and pre-trained on large-scale cerebrovascular disease related clinical text information. The pre-trained corpora contained 113,590 discharge notes, 105,743 radiology reports, and 38,199 neurological reports. Two real-world empirical clinical tasks were conducted to validate StrokeBERT's performance. The first task identified extracranial and intracranial artery stenosis from two independent sets of radiology angiography reports. The second task predicted the risk of recurrent ischemic stroke based on patients' first discharge information. RESULTS: In stenosis detection, StrokeBERT showed improved performance on targeted carotid arteries, with an average AUC compared to that of ClinicalBERT of 0.968 ± 0.021 and 0.956 ± 0.018, respectively. In recurrent ischemic stroke prediction, after 10-fold cross-validation on 1,700 discharge information, StrokeBERT presented better prediction ability (AUC±SD = 0.838 ± 0.017) than ClinicalBERT (AUC±SD = 0.808 ± 0.045). The attention scores of StrokeBERT showed better ability to detect and associate cerebrovascular disease related terms than current BERT based models. CONCLUSIONS: This study shows that a disease-specific BERT model improved the performance and accuracy of various disease-specific language processing tasks and can readily be fine-tuned to advance cerebrovascular disease research and further developed for clinical applications.

Year: 2021

Citations: 1

Full text: Yes

1. **Use of BERT (Bidirectional Encoder Representations from Transformers)-Based Deep Learning Method for Extracting Evidences in Chinese Radiology Reports: Development of a Computer-Aided Liver Cancer Diagnosis Framework.**

Abstract: BACKGROUND: Liver cancer is a substantial disease burden in China. As one of the primary diagnostic tools for detecting liver cancer, dynamic contrast-enhanced computed tomography provides detailed evidences for diagnosis that are recorded in free-text radiology reports. OBJECTIVE: The aim of our study was to apply a deep learning model and rule-based natural language processing (NLP) method to identify evidences for liver cancer diagnosis automatically. METHODS: We proposed a pretrained, fine-tuned BERT (Bidirectional Encoder Representations from Transformers)-based BiLSTM-CRF (Bidirectional Long Short-Term Memory-Conditional Random Field) model to recognize the phrases of APHE (hyperintense enhancement in the arterial phase) and PDPH (hypointense in the portal and delayed phases). To identify more essential diagnostic evidences, we used the traditional rule-based NLP methods for the extraction of radiological features. APHE, PDPH, and other extracted radiological features were used to design a computer-aided liver cancer diagnosis framework by random forest. RESULTS: The BERT-BiLSTM-CRF predicted the phrases of APHE and PDPH with an F1 score of 98.40% and 90.67%, respectively. The prediction model using combined features had a higher performance (F1 score, 88.55%) than those using APHE and PDPH (84.88%) or other extracted radiological features (83.52%). APHE and PDPH were the top 2 essential features for liver cancer diagnosis. CONCLUSIONS: This work was a comprehensive NLP study, wherein we identified evidences for the diagnosis of liver cancer from Chinese radiology reports, considering both clinical knowledge and radiology findings. The BERT-based deep learning method for the extraction of diagnostic evidence achieved state-of-the-art performance. The high performance proves the feasibility of the BERT-BiLSTM-CRF model in information extraction from Chinese radiology reports. The findings of our study suggest that the deep learning-based method for automatically identifying evidences for diagnosis can be extended to other types of Chinese clinical texts.

Year: 2021

Citations: 20

Full text: Yes

1. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**

Abstract: Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.

Year: 2019

Citations: 5.877

Full text: Yes

1. **Prediction of Chronic Kidney Disease Risk Using Multimodal Data**

Abstract: Chronic kidney disease (CKD) is a widespread public health problem and often leads to kidney failure which needs hemodialysis or even kidney transplantation. Undoubtedly, prediction of the risk of CKD among healthy people is highly desirable and very meaningful. However, most studies in this field used logistic regression (LR) and produced results with limited accuracy. Also, these studies ignored unstructured data which contained useful information. To improve CKD prediction, in this study, we built a novel multimodal data model that integrated Bidirectional Encoder Representations from Transformers with Light Gradient Boosting Machine (termed MD-BERT-LGBM model hereafter), and applied it to a group of 3295 participants for CKD prediction study. We collected medical data for over three months from each participant. We compared this novel integrated framework with three conventional models: the LR, LGBM, and Multimodal Disease Risk Prediction algorithm based on Convolutional Neural Networks (CNN-MDRP). The experimental results show that the new MD-BERT-LGBM model outperformed all the three conventional models in terms of accuracy, recall, and Area Under the ROC curve (AUC), which are 78.12%, 75.65%, and 85.15%, respectively. This result demonstrates the potential of this proposed method in the clinical application of CKD prediction and prevention.

Year: 2021

Citations: 1

Full text: Yes

1. **UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus**

Abstract: Contextual word embedding models, such as BioBERT and Bio\_ClinicalBERT, have achieved state-of-the-art results in biomedical natural language processing tasks by focusing their pre-training process on domain-specific corpora. However, such models do not take into consideration expert domain knowledge. In this work, we introduced UmlsBERT, a contextual embedding model that integrates domain knowledge during the pre-training process via a novel knowledge augmentation strategy. More specifically, the augmentation on UmlsBERT with the Unified Medical Language System (UMLS) Metathesaurus was performed in two ways: i) connecting words that have the same underlying `concept' in UMLS, and ii) leveraging semantic group knowledge in UMLS to create clinically meaningful input embeddings. By applying these two strategies, UmlsBERT can encode clinical domain knowledge into word embeddings and outperform existing domain-specific models on common named-entity recognition (NER) and clinical natural language inference clinical NLP tasks.

Year: 2020

Citations: 56

Full text: Yes

1. **A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study.**

Abstract: BACKGROUND: Information in pathology reports is critical for cancer care. Natural language processing (NLP) systems used to extract information from pathology reports are often narrow in scope or require extensive tuning. Consequently, there is growing interest in automated deep learning approaches. A powerful new NLP algorithm, bidirectional encoder representations from transformers (BERT), was published in late 2018. BERT set new performance standards on tasks as diverse as question answering, named entity recognition, speech recognition, and more. OBJECTIVE: The aim of this study is to develop a BERT-based system to automatically extract detailed tumor site and histology information from free-text oncological pathology reports. METHODS: We pursued three specific aims: extract accurate tumor site and histology descriptions from free-text pathology reports, accommodate the diverse terminology used to indicate the same pathology, and provide accurate standardized tumor site and histology codes for use by downstream applications. We first trained a base language model to comprehend the technical language in pathology reports. This involved unsupervised learning on a training corpus of 275,605 electronic pathology reports from 164,531 unique patients that included 121 million words. Next, we trained a question-and-answer (Q&A) model that connects a Q&A layer to the base pathology language model to answer pathology questions. Our Q&A system was designed to search for the answers to two predefined questions in each pathology report: What organ contains the tumor? and What is the kind of tumor or carcinoma? This involved supervised training on 8197 pathology reports, each with ground truth answers to these 2 questions determined by certified tumor registrars. The data set included 214 tumor sites and 193 histologies. The tumor site and histology phrases extracted by the Q&A model were used to predict International Classification of Diseases for Oncology, Third Edition (ICD-O-3), site and histology codes. This involved fine-tuning two additional BERT models: one to predict site codes and another to predict histology codes. Our final system includes a network of 3 BERT-based models. We call this CancerBERT network (caBERTnet). We evaluated caBERTnet using a sequestered test data set of 2050 pathology reports with ground truth answers determined by certified tumor registrars. RESULTS: caBERTnet's accuracies for predicting group-level site and histology codes were 93.53% (1895/2026) and 97.6% (1993/2042), respectively. The top 5 accuracies for predicting fine-grained ICD-O-3 site and histology codes with 5 or more samples each in the training data set were 92.95% (1794/1930) and 96.01% (1853/1930), respectively. CONCLUSIONS: We have developed an NLP system that outperforms existing algorithms at predicting ICD-O-3 codes across an extensive range of tumor sites and histologies. Our new system could help reduce treatment delays, increase enrollment in clinical trials of new therapies, and improve patient outcomes.

Year: 2022

Citations: 2

Full text: Yes

1. **SSN MLRG at ImageCLEFmedical Caption 2022: Medical Concept Detection and Caption Prediction using Transfer Learning and Transformer based Learning Approaches**

Abstract: The computer aided medical system for various applications is required now-a-days for an early and effective analysis. However most of the medical data are, publicly unavailable and exist in unstructured and unlabelled format are real challenges in developing the medical system. To address these issues, ImageCLEF forum is conducting many tasks on the medical domain from 2016 onwards. This year one of the tasks is medical concept detection and caption prediction. For this task, our team has proposed two concept detection techniques and caption prediction techniques. The concept detection models are developed using multi-label classification and information retrieval approaches resulted the F1-score and secondary F1score as 0.418 and 0.654 respectively. The caption prediction models are implemented using ResNet with Bidirectional Encoder Representations from Transformers (BERT) and, Sparse Auto Encoder (SAE) with Multi-Layer Perceptron (MLP) and Gated Recurrent Unit (GRU), which resulted a BLEU and BERT score of 0.160 and 0.545 respectively.

Year: 2022

Citations: -

Full text: Yes

1. **A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning.**

Abstract: BACKGROUND: Pathology synopses consist of semi-structured or unstructured text summarizing visual information by observing human tissue. Experts write and interpret these synopses with high domain-specific knowledge to extract tissue semantics and formulate a diagnosis in the context of ancillary testing and clinical information. The limited number of specialists available to interpret pathology synopses restricts the utility of the inherent information. Deep learning offers a tool for information extraction and automatic feature generation from complex datasets. METHODS: Using an active learning approach, we developed a set of semantic labels for bone marrow aspirate pathology synopses. We then trained a transformer-based deep-learning model to map these synopses to one or more semantic labels, and extracted learned embeddings (i.e., meaningful attributes) from the model's hidden layer. RESULTS: Here we demonstrate that with a small amount of training data, a transformer-based natural language model can extract embeddings from pathology synopses that capture diagnostically relevant information. On average, these embeddings can be used to generate semantic labels mapping patients to probable diagnostic groups with a micro-average F1 score of 0.779 Â ± 0.025. CONCLUSIONS: We provide a generalizable deep learning model and approach to unlock the semantic information inherent in pathology synopses toward improved diagnostics, biodiscovery and AI-assisted computational pathology.

Year: 2021

Citations: 1

Full text: Yes

1. **Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models**

Abstract: Clinical notes contain an extensive record of a patient's health status, such as smoking status or the presence of heart conditions. However, this detail is not replicated within the structured data of electronic health systems. Phenotyping, the extraction of patient conditions from free clinical text, is a critical task which supports avariety of downstream applications such as decision support and secondary use of medical records. Previous work has resulted in systems which are high performing but require hand engineering, often of rules. Recent work in pretrained contextualized language models have enabled advances in representing text for a variety of tasks. We therefore explore several architectures for modeling pheno-typing that rely solely on BERT representations of the clinical note, removing the need for manual engineering. We find these architectures are competitive with or outperform existing state of the art methods on two phenotyping tasks.

Year: 2019

Citations: 26

Full text: Yes

1. **Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT**

Abstract: Abstract Background The abundance of biomedical text data coupled with advances in natural language processing (NLP) is resulting in novel biomedical NLP (BioNLP) applications. These NLP applications, or tasks, are reliant on the availability of domain-specific language models (LMs) that are trained on a massive amount of data. Most of the existing domain-specific LMs adopted bidirectional encoder representations from transformers (BERT) architecture which has limitations, and their generalizability is unproven as there is an absence of baseline results among common BioNLP tasks. Results We present 8 variants of BioALBERT, a domain-specific adaptation of a lite bidirectional encoder representations from transformers (ALBERT), trained on biomedical (PubMed and PubMed Central) and clinical (MIMIC-III) corpora and fine-tuned for 6 different tasks across 20 benchmark datasets. Experiments show that a large variant of BioALBERT trained on PubMed outperforms the state-of-the-art on named-entity recognition (+ 11.09% BLURB score improvement), relation extraction (+ 0.80% BLURB score), sentence similarity (+ 1.05% BLURB score), document classification (+ 0.62% F1-score), and question answering (+ 2.83% BLURB score). It represents a new state-of-the-art in 5 out of 6 benchmark BioNLP tasks. Conclusions The large variant of BioALBERT trained on PubMed achieved a higher BLURB score than previous state-of-the-art models on 5 of the 6 benchmark BioNLP tasks. Depending on the task, 5 different variants of BioALBERT outperformed previous state-of-the-art models on 17 of the 20 benchmark datasets, showing that our model is robust and generalizable in the common BioNLP tasks. We have made BioALBERT freely available which will help the BioNLP community avoid computational cost of training and establish a new set of baselines for future efforts across a broad range of BioNLP tasks.

Year: 2022

Citations: 13

Full text: Yes

1. **Improved biomedical word embeddings in the transformer era.**

Abstract: BACKGROUND: Recent natural language processing (NLP) research is dominated by neural network methods that employ word embeddings as basic building blocks. Pre-training with neural methods that capture local and global distributional properties (e.g., skip-gram, GLoVE) using free text corpora is often used to embed both words and concepts. Pre-trained embeddings are typically leveraged in downstream tasks using various neural architectures that are designed to optimize task-specific objectives that might further tune such embeddings. OBJECTIVE: Despite advances in contextualized language model based embeddings, static word embeddings still form an essential starting point in BioNLP research and applications. They are useful in low resource settings and in lexical semantics studies. Our main goal is to build improved biomedical word embeddings and make them publicly available for downstream applications. METHODS: We jointly learn word and concept embeddings by first using the skip-gram method and further fine-tuning them with correlational information manifesting in co-occurring Medical Subject Heading (MeSH) concepts in biomedical citations. This fine-tuning is accomplished with the transformer-based BERT architecture in the two-sentence input mode with a classification objective that captures MeSH pair co-occurrence. We conduct evaluations of these tuned static embeddings using multiple datasets for word relatedness developed by previous efforts. RESULTS: Both in qualitative and quantitative evaluations we demonstrate that our methods produce improved biomedical embeddings in comparison with other static embedding efforts. Without selectively culling concepts and terms (as was pursued by previous efforts), we believe we offer the most exhaustive evaluation of biomedical embeddings to date with clear performance improvements across the board. CONCLUSION: We repurposed a transformer architecture (typically used to generate dynamic embeddings) to improve static biomedical word embeddings using concept correlations. We provide our code and embeddings for public use for downstream applications and research endeavors: <https://github.com/bionlproc/BERT-CRel-Embeddings>.

Year: 2021

Citations: 3

Full text: Yes

1. **The natural language processing of radiology requests and reports of chest imaging: Comparing five transformer models' multilabel classification and a proof-of-concept study**

Abstract: BACKGROUND: Radiology requests and reports contain valuable information about diagnostic findings and indications, and transformer-based language models are promising for more accurate text classification. METHODS: In a retrospective study, 2256 radiologist-annotated radiology requests (8 classes) and reports (10 classes) were divided into training and testing datasets (90% and 10%, respectively) and used to train 32 models. Performance metrics were compared by model type (LSTM, Bertje, RobBERT, BERT-clinical, BERT-multilingual, BERT-base), text length, data prevalence, and training strategy. The best models were used to predict the remaining 40,873 cases' categories of the datasets of requests and reports. RESULTS: The RobBERT model performed the best after 4000 training iterations, resulting in AUC values ranging from 0.808 [95% CI (0.757-0.859)] to 0.976 [95% CI (0.956-0.996)] for the requests and 0.746 [95% CI (0.689-0.802)] to 1.0 [95% CI (1.0-1.0)] for the reports. The AUC for the classification of normal reports was 0.95 [95% CI (0.922-0.979)]. The predicted data demonstrated variability of both diagnostic yield for various request classes and request patterns related to COVID-19 hospital admission data. CONCLUSION: Transformer-based natural language processing is feasible for the multilabel classification of chest imaging request and report items. Diagnostic yield varies with the information in the requests.

Year: 2022

Citations: -

Full text: Yes

1. **ConBERT: A Concatenation of Bidirectional Transformers for Standardization of Operative Reports from Electronic Medical Records**

Abstract: This operative report documents the details of a surgery. Standardization of the medical terminology for the operative report written in free text is significant for performing medical research and establishing insurance systems by accurately sharing information on treatment. However, standardization of operative reports is a labor-intensive task that has a risk of induced errors. We have proposed a concatenation of bidirectional encoder representations from transformers (ConBERT) model for predicting the International Classification of Disease-9 code using the operative report and diagnosis recorded in free text to standardize the operative report automatically. We compared the pre-trained models of BERT and character BERT and created a new model by concatenating the combinations of each model. The proposed ConBERT model showed a micro AP score of 0.7672, F1 score of 0.7415, and AUC of 0.9842. In addition, we developed a web-based application to demonstrate the performance of our model and make it publicly accessible.

Year: 2022

Citations: -

Full text: Yes

1. **Scalable and accurate deep learning with electronic health records**

Abstract: Abstract Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient’s record. We propose a representation of patients’ entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient’s final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient’s chart.

Year: 2018

Citations: 1.518

Full text: Yes

1. **Extracting medication changes in clinical narratives using pre-trained language models**

Abstract: An accurate and detailed account of patient medications, including medication changes within the patient timeline, is essential for healthcare providers to provide appropriate patient care. Healthcare providers or the patients themselves may initiate changes to patient medication. Medication changes take many forms, including prescribed medication and associated dosage modification. These changes provide information about the overall health of the patient and the rationale that led to the current care. Future care can then build on the resulting state of the patient. This work explores the automatic extraction of medication change information from free-text clinical notes. The Contextual Medication Event Dataset (CMED) is a corpus of clinical notes with annotations that characterize medication changes through multiple change-related attributes, including the type of change (start, stop, increase, etc.), initiator of the change, temporality, change likelihood, and negation. Using CMED, we identify medication mentions in clinical text and propose three novel high-performing BERT-based systems that resolve the annotated medication change characteristics. We demonstrate that our proposed systems improve medication change classification performance over the initial work exploring CMED.

Year: 2023

Citations: -

Full text: Yes

1. **Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction**

Abstract: Abstract Deep learning (DL)-based predictive models from electronic health records (EHRs) deliver impressive performance in many clinical tasks. Large training cohorts, however, are often required by these models to achieve high accuracy, hindering the adoption of DL-based models in scenarios with limited training data. Recently, bidirectional encoder representations from transformers (BERT) and related models have achieved tremendous successes in the natural language processing domain. The pretraining of BERT on a very large training corpus generates contextualized embeddings that can boost the performance of models trained on smaller datasets. Inspired by BERT, we propose Med-BERT, which adapts the BERT framework originally developed for the text domain to the structured EHR domain. Med-BERT is a contextualized embedding model pretrained on a structured EHR dataset of 28,490,650 patients. Fine-tuning experiments showed that Med-BERT substantially improves the prediction accuracy, boosting the area under the receiver operating characteristics curve (AUC) by 1.21–6.14% in two disease prediction tasks from two clinical databases. In particular, pretrained Med-BERT obtains promising performances on tasks with small fine-tuning training sets and can boost the AUC by more than 20% or obtain an AUC as high as a model trained on a training set ten times larger, compared with deep learning models without Med-BERT. We believe that Med-BERT will benefit disease prediction studies with small local training datasets, reduce data collection expenses, and accelerate the pace of artificial intelligence aided healthcare.

Year: 2021

Citations: 172

Full text: Yes

1. **Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models.**

Abstract: OBJECTIVE: A vast amount of medical data is still stored in unstructured text documents. We present an automated method of information extraction from German unstructured clinical routine data from the cardiology domain enabling their usage in state-of-the-art data-driven deep learning projects. METHODS: We evaluated pre-trained language models to extract a set of 12 cardiovascular concepts in German discharge letters. We compared three bidirectional encoder representations from transformers pre-trained on different corpora and fine-tuned them on the task of cardiovascular concept extraction using 204 discharge letters manually annotated by cardiologists at the University Hospital Heidelberg. We compared our results with traditional machine learning methods based on a long short-term memory network and a conditional random field. RESULTS: Our best performing model, based on publicly available German pre-trained bidirectional encoder representations from the transformer model, achieved a token-wise micro-average F1-score of 86% and outperformed the baseline by at least 6%. Moreover, this approach achieved the best trade-off between precision (positive predictive value) and recall (sensitivity). CONCLUSION: Our results show the applicability of state-of-the-art deep learning methods using pre-trained language models for the task of cardiovascular concept extraction using limited training data. This minimizes annotation efforts, which are currently the bottleneck of any application of data-driven deep learning projects in the clinical domain for German and many other European languages.

Year: 2021

Citations: 4

Full text: Yes

1. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**

Abstract: As Transfer Learning from large-scale pre-trained models becomes more prevalent in Natural Language Processing (NLP), operating these large models in on-the-edge and/or under constrained computational training or inference budgets remains challenging. In this work, we propose a method to pre-train a smaller general-purpose language representation model, called DistilBERT, which can then be fine-tuned with good performances on a wide range of tasks like its larger counterparts. While most prior work investigated the use of distillation for building task-specific models, we leverage knowledge distillation during the pre-training phase and show that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. To leverage the inductive biases learned by larger models during pre-training, we introduce a triple loss combining language modeling, distillation and cosine-distance losses. Our smaller, faster and lighter model is cheaper to pre-train and we demonstrate its capabilities for on-device computations in a proof-of-concept experiment and a comparative on-device study.

Year: 2019

Citations: 3.063

Full text: Yes

1. **Using deep learning-based natural language processing to identify reasons for statin nonuse in patients with atherosclerotic cardiovascular disease.**

Abstract: BACKGROUND: Statins conclusively decrease mortality in atherosclerotic cardiovascular disease (ASCVD), the leading cause of death worldwide, and are strongly recommended by guidelines. However, real-world statin utilization and persistence are low, resulting in excess mortality. Identifying reasons for statin nonuse at scale across health systems is crucial to developing targeted interventions to improve statin use. METHODS: We developed and validated deep learning-based natural language processing (NLP) approaches (Clinical Bidirectional Encoder Representations from Transformers [BERT]) to classify statin nonuse and reasons for statin nonuse using unstructured electronic health records (EHRs) from a diverse healthcare system. RESULTS: We present data from a cohort of 56,530 ASCVD patients, among whom 21,508 (38%) lack guideline-directed statin prescriptions and statins listed as allergies in structured EHR portions. Of these 21,508 patients without prescriptions, only 3,929 (18%) have any discussion of statin use or nonuse in EHR documentation. The NLP classifiers identify statin nonuse with an area under the curve (AUC) of 0.94 (95% CI 0.93-0.96) and reasons for nonuse with a weighted-average AUC of 0.88 (95% CI 0.86-0.91) when evaluated against manual expert chart review in a held-out test set. Clinical BERT identifies key patient-level reasons (side-effects, patient preference) and clinician-level reasons (guideline-discordant practices) for statin nonuse, including differences by type of ASCVD and patient race/ethnicity. CONCLUSIONS: Our deep learning NLP classifiers can identify crucial gaps in statin nonuse and reasons for nonuse in high-risk populations to support education, clinical decision support, and potential pathways for health systems to address ASCVD treatment gaps.

Year: 2022

Citations: 1

Full text: Yes

1. **Learning unsupervised contextual representations for medical synonym discovery**

Abstract: Abstract Objectives An important component of processing medical texts is the identification of synonymous words or phrases. Synonyms can inform learned representations of patients or improve linking mentioned concepts to medical ontologies. However, medical synonyms can be lexically similar (“dilated RA” and “dilated RV”) or dissimilar (“cerebrovascular accident” and “stroke”); contextual information can determine if 2 strings are synonymous. Medical professionals utilize extensive variation of medical terminology, often not evidenced in structured medical resources. Therefore, the ability to discover synonyms, especially without reliance on training data, is an important component in processing training notes. The ability to discover synonyms from models trained on large amounts of unannotated data removes the need to rely on annotated pairs of similar words. Models relying solely on non-annotated data can be trained on a wider variety of texts without the cost of annotation, and thus may capture a broader variety of language. Materials and Methods Recent contextualized deep learning representation models, such as ELMo (Peters et al., 2019) and BERT, (Devlin et al. 2019) have shown strong improvements over previous approaches in a broad variety of tasks. We leverage these contextualized deep learning models to build representations of synonyms, which integrate the context of surrounding sentence and use character-level models to alleviate out-of-vocabulary issues. Using these models, we perform unsupervised discovery of likely synonym matches, which reduces the reliance on expensive training data. Results We use the ShARe/CLEF eHealth Evaluation Lab 2013 Task 1b data to evaluate our synonym discovery method. Comparing our proposed contextualized deep learning representations to previous non-neural representations, we find that the contextualized representations show consistent improvement over non-contextualized models in all metrics. Conclusions Our results show that contextualized models produce effective representations for synonym discovery. We expect that the use of these representations in other tasks would produce similar gains in performance.

Year: 2019

Citations: 11

Full text: Yes

1. **Three-level Hierarchical Transformer Networks for Long-sequence and Multiple Clinical Documents Classification**

Abstract: We present a Three-level Hierarchical Transformer Network (3-level-HTN) for modeling long-term dependencies across clinical notes for the purpose of patient-level prediction. The network is equipped with three levels of Transformer-based encoders to learn progressively from words to sentences, sentences to notes, and finally notes to patients. The first level from word to sentence directly applies a pre-trained BERT model as a fully trainable component. While the second and third levels both implement a stack of transformer-based encoders, before the final patient representation is fed into a classification layer for clinical predictions. Compared to conventional BERT models, our model increases the maximum input length from 512 tokens to much longer sequences that are appropriate for modeling large numbers of clinical notes. We empirically examine different hyper-parameters to identify an optimal trade-off given computational resource limits. Our experiment results on the MIMIC-III dataset for different prediction tasks demonstrate that the proposed Hierarchical Transformer Network outperforms previous state-of-the-art models, including but not limited to BigBird.

Year: 2021

Citations: -

Full text: Yes

1. **Exploring Language Markers of Mental Health in Psychiatric Stories**

Abstract: Diagnosing mental disorders is complex due to the genetic, environmental and psychological contributors and the individual risk factors. Language markers for mental disorders can help to diagnose a person. Research thus far on language markers and the associated mental disorders has been done mainly with the Linguistic Inquiry and Word Count (LIWC) program. In order to improve on this research, we employed a range of Natural Language Processing (NLP) techniques using LIWC, spaCy, fastText and RobBERT to analyse Dutch psychiatric interview transcriptions with both rule-based and vector-based approaches. Our primary objective was to predict whether a patient had been diagnosed with a mental disorder, and if so, the specific mental disorder type. Furthermore, the second goal of this research was to find out which words are language markers for which mental disorder. LIWC in combination with the random forest classification algorithm performed best in predicting whether a person had a mental disorder or not (accuracy: 0.952; Cohen’s kappa: 0.889). SpaCy in combination with random forest predicted best which particular mental disorder a patient had been diagnosed with (accuracy: 0.429; Cohen’s kappa: 0.304).

Year: 2022

Citations: 8

Full text:

1. **TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning**

Abstract: Masked language models (MLMs) such as BERT and RoBERTa have revolutionized the field of Natural Language Understanding in the past few years. However, existing pre-trained MLMs often output an anisotropic distribution of token representations that occupies a narrow subset of the entire representation space. Such token representations are not ideal, especially for tasks that demand discriminative semantic meanings of distinct tokens. In this work, we propose TaCL (Token-aware Contrastive Learning), a novel continual pre-training approach that encourages BERT to learn an isotropic and discriminative distribution of token representations. TaCL is fully unsupervised and requires no additional data. We extensively test our approach on a wide range of English and Chinese benchmarks. The results show that TaCL brings consistent and notable improvements over the original BERT model. Furthermore, we conduct detailed analysis to reveal the merits and inner-workings of our approach.

Year: 2021

Citations: 20

Full text: Yes

1. **How to Fine-Tune BERT for Text Classification?**

Abstract: Language model pre-training has proven to be useful in learning universal language representations. As a state-of-the-art language model pre-training model, BERT (Bidirectional Encoder Representations from Transformers) has achieved amazing results in many language understanding tasks. In this paper, we conduct exhaustive experiments to investigate different fine-tuning methods of BERT on text classification task and provide a general solution for BERT fine-tuning. Finally, the proposed solution obtains new state-of-the-art results on eight widely-studied text classification datasets.

Year: 2019

Citations: 1.105

Full text: Yes

1. **Identifying Patient Smoking Status from Medical Discharge Records**

Abstract: -

Year: 2008

Citations: 416

Full text: Yes

1. **Negation detection in Dutch clinical texts: an evaluation of rule-based and machine learning methods**

Abstract: As structured data are often insufficient, labels need to be extracted from free text in electronic health records when developing models for clinical information retrieval and decision support systems. One of the most important contextual properties in clinical text is negation, which indicates the absence of findings. We aimed to improve large scale extraction of labels by comparing three methods for negation detection in Dutch clinical notes. We used the Erasmus Medical Center Dutch Clinical Corpus to compare a rule-based method based on ContextD, a biLSTM model using MedCAT and (finetuned) RoBERTa-based models. We found that both the biLSTM and RoBERTa models consistently outperform the rule-based model in terms of F1 score, precision and recall. In addition, we systematically categorized the classification errors for each model, which can be used to further improve model performance in particular applications. Combining the three models naively was not beneficial in terms of performance. We conclude that the biLSTM and RoBERTa-based models in particular are highly accurate accurate in detecting clinical negations, but that ultimately all three approaches can be viable depending on the use case at hand.

Year: 2023

Citations: -

Full text: Yes

1. **Predicting COVID-19 Symptoms From Free Text in Medical Records Using Artificial Intelligence: Feasibility Study**

Abstract: Background Electronic medical records have opened opportunities to analyze clinical practice at large scale. Structured registries and coding procedures such as the International Classification of Primary Care further improved these procedures. However, a large part of the information about the state of patient and the doctors’ observations is still entered in free text fields. The main function of those fields is to report the doctor’s line of thought, to remind oneself and his or her colleagues on follow-up actions, and to be accountable for clinical decisions. These fields contain rich information that can be complementary to that in coded fields, and until now, they have been hardly used for analysis. Objective This study aims to develop a prediction model to convert the free text information on COVID-19–related symptoms from out of hours care electronic medical records into usable symptom-based data that can be analyzed at large scale. Methods The design was a feasibility study in which we examined the content of the raw data, steps and methods for modelling, as well as the precision and accuracy of the models. A data prediction model for 27 preidentified COVID-19–relevant symptoms was developed for a data set derived from the database of primary-care out-of-hours consultations in Flanders. A multiclass, multilabel categorization classifier was developed. We tested two approaches, which were (1) a classical machine learning–based text categorization approach, Binary Relevance, and (2) a deep neural network learning approach with BERTje, including a domain-adapted version. Ethical approval was acquired through the Institutional Review Board of the Institute of Tropical Medicine and the ethics committee of the University Hospital of Antwerpen (ref 20/50/693). Results The sample set comprised 3957 fields. After cleaning, 2313 could be used for the experiments. Of the 2313 fields, 85% (n=1966) were used to train the model, and 15% (n=347) for testing. The normal BERTje model performed the best on the data. It reached a weighted F1 score of 0.70 and an exact match ratio or accuracy score of 0.38, indicating the instances for which the model has identified all correct codes. The other models achieved respectable results as well, ranging from 0.59 to 0.70 weighted F1. The Binary Relevance method performed the best on the data without a frequency threshold. As for the individual codes, the domain-adapted version of BERTje performs better on several of the less common objective codes, while BERTje reaches higher F1 scores for the least common labels especially, and for most other codes in general. Conclusions The artificial intelligence model BERTje can reliably predict COVID-19–related information from medical records using text mining from the free text fields generated in primary care settings. This feasibility study invites researchers to examine further possibilities to use primary care routine data.

Year: 2022

Citations: 2

Full text: Yes

1. **MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records**

Abstract: This paper presents MedRoBERTa.nl as the first Transformer-based language model for Dutch medical language. We show that using 13GB of text data from Dutch hospital notes, pre-training from scratch results in a better domain-specific language model than further pre-training RobBERT. When extending pre-training on RobBERT, we use a domain-specific vocabulary and re-train the embedding look-up layer. We show that MedRoBERTa.nl, the model that was trained from scratch, outperforms general language models for Dutch on a medical odd-one-out similarity task. MedRoBERTa.nl already reaches higher performance than general language models for Dutch on this task after only 10k pre-training steps. When fine-tuned, MedRobERTa.nl outperforms general language models for Dutch in a task classifying sentences from Dutch hospital notes that contain information about patients’ mobility levels.

Year: 2021

Citations: 4

Full text: Yes

1. **Pre-training technique to localize medical BERT and enhance biomedical BERT**

Abstract: Pre-training large-scale neural language models on raw texts has made a significant contribution to improving transfer learning in natural language processing (NLP). With the introduction of transformer-based language models, such as bidirectional encoder representations from transformers (BERT), the performance of information extraction from a free text by NLP has significantly improved for both the general domain and medical domain; however, it is difficult to train specific BERT models that perform well for domains in which there are few publicly available databases of high quality and large size. We hypothesized that this problem can be addressed by up-sampling a domain-specific corpus and using it for pre-training with a larger corpus in a balanced manner. Our proposed method consists of a single intervention with one option: simultaneous pre-training after up-sampling and amplified vocabulary. We conducted three experiments and evaluated the resulting products. We confirmed that our Japanese medical BERT outperformed conventional baselines and the other BERT models in terms of the medical document classification task and that our English BERT pre-trained using both the general and medical-domain corpora performed sufficiently well for practical use in terms of the biomedical language understanding evaluation (BLUE) benchmark. Moreover, our enhanced biomedical BERT model, in which clinical notes were not used during pre-training, showed that both the clinical and biomedical scores of the BLUE benchmark were 0.3 points above that of the ablation model trained without our proposed method. Well-balanced pre-training by up-sampling instances derived from a corpus appropriate for the target task allows us to construct a high-performance BERT model.

Year: 2020

Citations: 11

Full text: Yes

1. **Dutch Humor Detection by Generating Negative Examples**

Abstract: Detecting if a text is humorous is a hard task to do computationally, as it usually requires linguistic and common sense insights. In machine learning, humor detection is usually modeled as a binary classification task, trained to predict if the given text is a joke or another type of text. Rather than using completely different non-humorous texts, we propose using text generation algorithms for imitating the original joke dataset to increase the difficulty for the learning algorithm. We constructed several different joke and non-joke datasets to test the humor detection abilities of different language technologies. In particular, we compare the humor detection capabilities of classic neural network approaches with the state-of-the-art Dutch language model RobBERT. In doing so, we create and compare the first Dutch humor detection systems. We found that while other language models perform well when the non-jokes came from completely different domains, RobBERT was the only one that was able to distinguish jokes from generated negative examples. This performance illustrates the usefulness of using text generation to create negative datasets for humor recognition, and also shows that transformer models are a large step forward in humor detection.

Year: 2020

Citations: 7

Full text: Yes

1. **Text-based classification of interviews for mental health -- juxtaposing the state of the art**

Abstract: Currently, the state of the art for classification of psychiatric illness is based on audio-based classification. This thesis aims to design and evaluate a state of the art text classification network on this challenge. The hypothesis is that a well designed text-based approach poses a strong competition against the state-of-the-art audio based approaches. Dutch natural language models are being limited by the scarcity of pre-trained monolingual NLP models, as a result Dutch natural language models have a low capture of long range semantic dependencies over sentences. For this issue, this thesis presents belabBERT, a new Dutch language model extending the RoBERTa[15] architecture. belabBERT is trained on a large Dutch corpus (+32GB) of web crawled texts. After this thesis evaluates the strength of text-based classification, a brief exploration is done, extending the framework to a hybrid text- and audio-based classification. The goal of this hybrid framework is to show the principle of hybridisation with a very basic audio-classification network. The overall goal is to create the foundations for a hybrid psychiatric illness classification, by proving that the new text-based classification is already a strong stand-alone solution.

Year: 2020

Citations: 1

Full text: Yes

1. **Research on Named Entity Recognition of Electronic Medical Records Based on RoBERTa and Radical-Level Feature**

Abstract: Clinical named entity recognition (CNER) identifies entities from unstructured medical records and classifies them into predefined categories. It is of great significance for follow-up clinical studies. Most of the existing CNER methods fail to give enough thought to Chinese radical-level characteristics and the specialty of the Chinese field. This paper proposes the Ra-RC model, which combines radical features and a deep learning structure to fix this problem. A bidirectional encoder representation of transformer (RoBERTa) is utilized to learn medical features thoroughly. Simultaneously, we use the bidirectional long short-term memory (BiLSTM) network to extract radical-level information to capture the internal relevance of characteristics and stitch the eigenvectors generated by RoBERTa. In addition, the relationship between labels is considered to obtain the optimal tag sequence by applying conditional random field (CRF). The experimental results demonstrate that the proposed Ra-RC model achieves F1 score 93.26% and 82.87% on the CCKS2017 and CCKS2019 datasets, respectively.

Year: 2021

Citations: 6

Full text: Yes

1. **Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing.**

Abstract: OBJECTIVE: Seizure frequency and seizure freedom are among the most important outcome measures for patients with epilepsy. In this study, we aimed to automatically extract this clinical information from unstructured text in clinical notes. If successful, this could improve clinical decision-making in epilepsy patients and allow for rapid, large-scale retrospective research. MATERIALS AND METHODS: We developed a finetuning pipeline for pretrained neural models to classify patients as being seizure-free and to extract text containing their seizure frequency and date of last seizure from clinical notes. We annotated 1000 notes for use as training and testing data and determined how well 3 pretrained neural models, BERT, RoBERTa, and Bio\_ClinicalBERT, could identify and extract the desired information after finetuning. RESULTS: The finetuned models (BERTFT, Bio\_ClinicalBERTFT, and RoBERTaFT) achieved near-human performance when classifying patients as seizure free, with BERTFT and Bio\_ClinicalBERTFT achieving accuracy scores over 80%. All 3 models also achieved human performance when extracting seizure frequency and date of last seizure, with overall F1 scores over 0.80. The best combination of models was Bio\_ClinicalBERTFT for classification, and RoBERTaFT for text extraction. Most of the gains in performance due to finetuning required roughly 70 annotated notes. DISCUSSION AND CONCLUSION: Our novel machine reading approach to extracting important clinical outcomes performed at or near human performance on several tasks. This approach opens new possibilities to support clinical practice and conduct large-scale retrospective clinical research. Future studies can use our finetuning pipeline with minimal training annotations to answer new clinical questions.

Year: 2022

Citations: 7

Full text: Yes

1. **Traditional Chinese medicine clinical records classification with BERT and domain specific corpora**

Abstract: Abstract Traditional Chinese Medicine (TCM) has been developed for several thousand years and plays a significant role in health care for Chinese people. This paper studies the problem of classifying TCM clinical records into 5 main disease categories in TCM. We explored a number of state-of-the-art deep learning models and found that the recent Bidirectional Encoder Representations from Transformers can achieve better results than other deep learning models and other state-of-the-art methods. We further utilized an unlabeled clinical corpus to fine-tune the BERT language model before training the text classifier. The method only uses Chinese characters in clinical text as input without preprocessing or feature engineering. We evaluated deep learning models and traditional text classifiers on a benchmark data set. Our method achieves a state-of-the-art accuracy 89.39% ± 0.35%, Macro F1 score 88.64% ± 0.40% and Micro F1 score 89.39% ± 0.35%. We also visualized attention weights in our method, which can reveal indicative characters in clinical text.

Year: 2019

Citations: 37

Full text: Yes

1. **A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models.**

Abstract: Social and behavioral determinants of health (SBDoH) have important roles in shaping people's health. In clinical research studies, especially comparative effectiveness studies, failure to adjust for SBDoH factors will potentially cause confounding issues and misclassification errors in either statistical analyses and machine learning-based models. However, there are limited studies to examine SBDoH factors in clinical outcomes due to the lack of structured SBDoH information in current electronic health record (EHR) systems, while much of the SBDoH information is documented in clinical narratives. Natural language processing (NLP) is thus the key technology to extract such information from unstructured clinical text. However, there is not a mature clinical NLP system focusing on SBDoH. In this study, we examined two state-of-the-art transformer-based NLP models, including BERT and RoBERTa, to extract SBDoH concepts from clinical narratives, applied the best performing model to extract SBDoH concepts on a lung cancer screening patient cohort, and examined the difference of SBDoH information between NLP extracted results and structured EHRs (SBDoH information captured in standard vocabularies such as the International Classification of Diseases codes). The experimental results show that the BERT-based NLP model achieved the best strict/lenient F1-score of 0.8791 and 0.8999, respectively. The comparison between NLP extracted SBDoH information and structured EHRs in the lung cancer patient cohort of 864 patients with 161,933 various types of clinical notes showed that much more detailed information about smoking, education, and employment were only captured in clinical narratives and that it is necessary to use both clinical narratives and structured EHRs to construct a more complete picture of patients' SBDoH factors.

Year: 2021

Citations: 4

Full text: Yes

1. **Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods.**

Abstract: BACKGROUND: Diabetic retinopathy (DR) is a leading cause of blindness in American adults. If detected, DR can be treated to prevent further damage causing blindness. There is an increasing interest in developing artificial intelligence (AI) technologies to help detect DR using electronic health records. The lesion-related information documented in fundus image reports is a valuable resource that could help diagnoses of DR in clinical decision support systems. However, most studies for AI-based DR diagnoses are mainly based on medical images; there is limited studies to explore the lesion-related information captured in the free text image reports. METHODS: In this study, we examined two state-of-the-art transformer-based natural language processing (NLP) models, including BERT and RoBERTa, compared them with a recurrent neural network implemented using Long short-term memory (LSTM) to extract DR-related concepts from clinical narratives. We identified four different categories of DR-related clinical concepts including lesions, eye parts, laterality, and severity, developed annotation guidelines, annotated a DR-corpus of 536 image reports, and developed transformer-based NLP models for clinical concept extraction and relation extraction. We also examined the relation extraction under two settings including 'gold-standard' setting-where gold-standard concepts were used-and end-to-end setting. RESULTS: For concept extraction, the BERT model pretrained with the MIMIC III dataset achieve the best performance (0.9503 and 0.9645 for strict/lenient evaluation). For relation extraction, BERT model pretrained using general English text achieved the best strict/lenient F1-score of 0.9316. The end-to-end system, BERT\_general\_e2e, achieved the best strict/lenient F1-score of 0.8578 and 0.8881, respectively. Another end-to-end system based on the RoBERTa architecture, RoBERTa\_general\_e2e, also achieved the same performance as BERT\_general\_e2e in strict scores. CONCLUSIONS: This study demonstrated the efficiency of transformer-based NLP models for clinical concept extraction and relation extraction. Our results show that it's necessary to pretrain transformer models using clinical text to optimize the performance for clinical concept extraction. Whereas, for relation extraction, transformers pretrained using general English text perform better.

Year: 2022

Citations: -

Full text: Yes

1. **A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model.**

Abstract: With the current advancement in the Internet, there has been a growing demand for building intelligent and smart systems that can efficiently address the detection of health-related problems on social media, such as the detection of depression and anxiety. These types of systems, which are mainly dependent on machine learning techniques, must be able to deal with obtaining the semantic and syntactic meaning of texts posted by users on social media. The data generated by users on social media contains unstructured and unpredictable content. Several systems based on machine learning and social media platforms have recently been introduced to identify health-related problems. However, the text representation and deep learning techniques employed provide only limited information and knowledge about the different texts posted by users. This is owing to a lack of long-term dependencies between each word in the entire text and a lack of proper exploitation of recent deep learning schemes. In this paper, we propose a novel framework to efficiently and effectively identify depression and anxiety-related posts while maintaining the contextual and semantic meaning of the words used in the whole corpus when applying bidirectional encoder representations from transformers (BERT). In addition, we propose a knowledge distillation technique, which is a recent technique for transferring knowledge from a large pretrained model (BERT) to a smaller model to boost performance and accuracy. We also devised our own data collection framework from Reddit and Twitter, which are the most common social media sites. Finally, we employed word2vec and BERT with Bi-LSTM to effectively analyze and detect depression and anxiety signs from social media posts. Our system surpasses other state-of-the-art methods and achieves an accuracy of 98% using the knowledge distillation technique.

Year: 2022

Citations: 10

Full text: Yes

1. **Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study.**

Abstract: BACKGROUND: Family history information, including information on family members, side of the family of family members, living status of family members, and observations of family members, plays an important role in disease diagnosis and treatment. Family member information extraction aims to extract family history information from semistructured/unstructured text in electronic health records (EHRs), which is a challenging task regarding named entity recognition (NER) and relation extraction (RE), where named entities refer to family members, living status, and observations, and relations refer to relations between family members and living status, and relations between family members and observations. OBJECTIVE: This study aimed to introduce the system we developed for the 2019 n2c2/OHNLP track on family history extraction, which can jointly extract entities and relations about family history information from clinical text. METHODS: We proposed a novel graph-based model with biaffine attention for family history extraction from clinical text. In this model, we first designed a graph to represent family history information, that is, representing NER and RE regarding family history in a unified way, and then introduced a biaffine attention mechanism to extract family history information in clinical text. Convolution neural network (CNN)-Bidirectional Long Short Term Memory network (BiLSTM) and Bidirectional Encoder Representation from Transformers (BERT) were used to encode the input sentence, and a biaffine classifier was used to extract family history information. In addition, we developed a postprocessing module to adjust the results. A system based on the proposed method was developed for the 2019 n2c2/OHNLP shared task track on family history information extraction. RESULTS: Our system ranked first in the challenge, and the F1 scores of the best system on the NER subtask and RE subtask were 0.8745 and 0.6810, respectively. After the challenge, we further fine tuned the parameters and improved the F1 scores of the two subtasks to 0.8823 and 0.7048, respectively. CONCLUSIONS: The experimental results showed that the system based on the proposed method can extract family history information from clinical text effectively.

Year: 2021

Citations: 1

Full text: Yes

# Classification table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Name paper** | **Type of task** | **Name of (BERT) model(s) used** | **Methods** | **Type and amount of training data** | **Results** |
| 2 | Identification of asthma control factor  in clinical notes using a hybrid deep learning  model | Identifying the description of reviewing inhaler technique in clinical notes. Comparison of several BERT models with several different setups. | BERT and cBERT (clinical bert) | Training BERT and cBERT using weakly-labelled data. Applying post-hoc rules to fix trivial errors. Comparing a rule-based model, standard BERT and cBERT, BERT and cBERT with distant supervision and BERT and cBERT with a hybrid approach of post-hoc rules were all compared to each other. Addition of dropout layer and linear classification layer. Application of cost-sensitivity to deal with imbalanced data. Usage of cyclical learning rate with triangular mode scheduler. Application of post-hoc rules to fix errors from BERT. | Free text clinical notes. 1.039 clinical notes of 300 patients that were manually chart reviewed. 27.363 clinical notes of 800 patients which were labelled weakly. The data has a severe class imbalance (only 0.124% of the input data was positive) | F1-scores: cBERT with post-hoc rules and distant supervision: **0.904**, BERT with post-hoc rules and distant supervision: 0.877, cBERT with distant supervision: 0.88, BERT with distant supervision: 0.853, cBERT: 0.838, BERT: 0.845, just rules: 0.837. |
| Comment: This paper shows different setups for BERT and clinical BERT (cBERT), whilst also showing which one performs better in comparison to the others on a classification task. The fact that this model was trained on input data with severe class imbalance and still manages to perform greatly is a big plus. It could be interesting to apply and compare the same structures on my input data, for example for drug users as that is almost never mentioned in a clinical note. | | | | | | |
| **ID** | **Name paper** | **Type of task** | **Name of (BERT) model(s) used** | **Methods** | **Type and amount of training data** | **Results** |
| 3 | Analysis of Language Embeddings for Classification of Unstructured Pathology Reports | Extracting meaningful embeddings from written pathology reports with the goal of classifying various types of cancer, comparison BERT models and TF-IDF | BioBERT, Clinical BioBERT and BioBERT in combination with RoBERTa | Data is preprocessed with a maximum length of 300 per report. Preprocessing includes removing stop words and standardization. The BERT models are used to tokenize the data. Neural network with two separate branches, one for the BERT models and one for the TF IDF model. The word embeddings obtained from the BERT models are passed through a Bidirectional LSTM layer and are then concatenated to the TF-IDF feature vectors and then passed through a dense and dropout layer. Ultimately, the final layer is a dense output layer with a softmax activation. Results are evaluated using 5-fold cross validation. Concatenation with TF-IDF feature vectors. | 1.960 free text pathology reports describing the tissues of organs referred to as the disease type from the publicly available Cancer Genome Atlas. Each report has 0 to 2.500 words. The disease type consists of seven classes. | F1-scores: Clinical BioBERT model that makes use of contextualized word embeddings along with TF-IDF feature vectors:  **0.91**, BioBERT with TF-IDF: 0.88, BioMed-RoBERTa with TF-IDF: 0.9, TF-IDF alone: 0.81, BioMed-RoBERTa alone: 0.83. |
| Comment: This paper shows the value of adding TF-IDF feature vectors to a BERT model and shows that it can greatly increase performance, as its addition caused an increase in F1-score of more than 12%. It is unknown how well this setup would perform on larger and more diverse data, as I believe TF-IDF would be less viable on a way larger dataset. It could be worth trying if this setup is used in other sources as well. | | | | | | |
| **ID** | **Name paper** | **Type of task** | **Name of (BERT) model(s) used** | **Methods** | **Type and amount of training data** | **Results** |
| 5 | MLT-DFKI at CLEF eHealth 2019:  Multi-label Classification of ICD-10 Codes with  BERT | Classifying International Statistical Classification of Diseases (ICD) codes from free text of animal experimentation of disease cure research. Multi-label classification as documents can have more than one code. | BERT and BioBERT | German input texts from the MLT-DFKI task are translated to English. The baseline is a TF-IDF weighted bag-of-words based linear SVM model. This model is compared to a CNN, attention models, HAN, SLSTM, CLSTM and BERT. The task is classification of ICD-10 codes in German non-technical summaries of animal experiments. Standard BERT model is used but with a max sequence length of 256 and a batch size of 6. | 8.385 training documents and 407 test documents, all in German. Each document has six text fields corresponding to title, goals, possible harms caused to animals and comments. Each document is assigned one or more ICD-10 codes. | Translating the text to English improved the score by an average of about 4%. On the test set, BERT models outperformed the other models in both German and English, with English BioBERT performing slightly better than English BERT. The BERT models in general had far superior recall but slightly inferior precision. Finally, the English BioBERT model and a CLSTSM model were combined to get an additional F1-score increase of almost 5%. This ensemble model ended up with an F1-score of 0.7798. As the task was for single model submissions only they handed in their BioBERT model which achieved an F1-score of 0.7302. |
| Comments: I do not believe this paper is useful for me, for a multitude of reasons. Firstly, it uses the generic BERT models and even decreases its power, causing a very noticeable decrease in performance. Secondly, I think the ultimate F1-scores are relatively low compared to results achieved by other papers, especially because of the hyperspecific training task, where one would expect higher scores. The paper does pose an interesting question however, which is the one of translating the texts to English. As we are dealing with Dutch texts, translating them to English might be an essential part of obtaining acceptable performance in BERT models trained on English input such as BERT and BioBERT respectively. This paper shows that translation could lead to a significant increase in performance. | | | | | | |
| **ID** | **Name paper** | **Type of task** | **Name of (BERT) model(s) used** | **Methods** | **Type and amount of training data** | **Results** |
| 6 | Classification of Medical Image Notes for Image Labeling  by Using MinBERT | Classifying patient image note data, in image type identification and clinical diagnosis identification. | BioBERT, ClinicalBERT, MinBERT. | Extracting useful label information for medical image classification. Use of MinBERT as it has been shown to increase performance. Compared to Word2Vec and CNN among others as baselines. BioBERT is also used as comparison method. Data was preprocessed by selecting six types of entities synthetically, on which the classification took place. | 15.466 valid instances of medical image notes, selected from a total 282.740. 80% of the set was used as training and the remaining 20% was used as test set. | F1-scores average on five types of image groups: MinBERT: **0.9952**, Word2Vec + CNN: 0.9947, ClinicalBERT: 0.966, BioBERT: 0.9543 |
| Comments: This study shows that a high performance is feasible on a smaller subset of a large set of medical image notes. As this study focuses on a set that is smaller than ours (just image notes rather than all clinical notes) and makes further distinctions by selecting a smaller subset on a hyperspecific task. It is hard to figure out how well this method extends to larger input sets. | | | | | | |
| 16 | Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning | Classifying International Statistical Classification of Diseases (ICD) codes from free text medical notes. | BERT, ClinicalBERT, BioBERT | Use of models word2vec, GLoVE, ELMo and SHARNN as well as BERT models specified in previous column. Preprocessing was applied by removing Chinese words and stop words among other measures. A classification model was obtained by constructing a model with 4 neural network layers, including a word embedding layer, a BiGRU layer, the parameters of which are the same for each pretrained model. The output layer is set to a dimension of 14.602 by 9.780 to correspond to the amount of different labels there are. | Data was acquired from patients at the National Taiwan University Hospital, from the paper alone it is unknown how many records there are in total. The ICD-10 codes are manually annotated. | F1-scores of whole label classification: word2vec: 0.68, GloVE: 0.635, ELMo: 0.631, BERT: 0.71, ClinicalBERT: **0.714**, BioBERT: 0.701, SHARNN: 0.57 |
| Comments: This paper is interesting due to the many possible labels in its multi-label classification task. However, as the size of the dataset cannot be inferred from the paper it is hard to place its results in the proper context. This also makes this paper hard to compare to other papers, which leads me to not consider its contents further. | | | | | | |
| 18 | Transformer-based models for ICD-10 coding of death certificates with Portuguese text | Proposing a BERT model to assign ICD-10 codes for cause of death to free-text descriptions in death certificates. Multi-label classification. | BERT, Portuguese BERTlarge | For pretraining, a novel process is proposed which incorporates in-domain knowledge. The pretraining phase consists of using the original MLM task but replacing the NSP with another task that better resembles the final task of automatic ICD coding. This new task has the objective to identify whether a death certificate is real or artificially created. Finetuning is done by adding a fully-connected classification output layer to the BERT model. | The data consists of Portuguese death certificates. A total of 121.536 instances were considered, where 75% was used for training and 25% for testing. | ICD-10 chapter classification F1-scores on test set: baseline of deep neural network of word embeddings: 0.629, BERT: 0.683, Finetuned BERT model pretrained on source data with MLM and novel clinical pre-training method: **0.723**. |
| Comments: This paper shows an alternative to next sentence prediction as a pretraining objective which is some sort of discriminator task between real and fake death certificates. The multi-class classification aspect of the task at hand, combined with the achieved increase in performance could warrant me trying this other pretraining method, although there are no results shown of a BERT model that was pretrained on the source data and made use of the original MLM and NSP tasks, which makes it difficult to determine what the effect of using the new pretraining objective really is. | | | | | | |
| 19 | Multi-label Classification for Clinical Text with Feature-level Attention | Training a new BERT model (FAMLC-BERT) that makes use of feature-level attention with the goal of improving the field of multi-labelled disease classification on Electronic Health Records (EHR). | FAMLC-BERT, BERT | Multi-label text classification is applied to predict diagnoses from clinical plain text. The model measures the probability of each of the disease labels. BERT is used to construct a distributed representation of each sentence in the free text. Feature-level attention is used to aggregate features that have different semantic information from different encoder layers. All hidden vectors are aggregated to apply attention at feature-level. Performance is compared to CNN, fastText, Text-RNN, HAN, CRNN and regular BERT models. | Total 6.435 EHRs were obtained from hospitals. For every record there are eleven implicit symptoms as well as labels. The data is split into training and test set in a 70-30 ratio. | F1-score on test set (k=1): Text-CNN: 0.4684, fasText: 0.4655, XML-CNN: 0.5083, BERT: 0.6129, FAMLC-BERT: **0.7907** |
| Comments: This paper shows a rather unique approach with the goal of capturing deep semantic information and shows that it can be feasible on a multi-label classification task. It also claims to capture context even when words indicating context are far apart in the document. The drastic increase in performance shows that training my model like this could be useful for my classification task as well, especially because it is also a multi-class classification problem. | | | | | | |
| 25 | Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text | Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text | BioClinicalBERT (small, medium and large), BioBERT Pubmed, BioClinicalBERT COVID, SciBERT | Models were trained on X-ray and CT-scan reports. Six different input datasets concerning X-ray and/or CT scan text reports were used for the classification task, which is a Neer classification. Neer divides the proximal humerus into seven parts. As entries can belong to more than one class this is a multi-class classification problem. A multitude of models get trained on this task and their performance is measured. These models are statistical ML models like Multinomial Naïve Bayes, Random Forest and SGD and BERT models specified in the previous column. In order to battle class imbalance three oversampling techniques were applied. BERT models were trained for 10 epochs. | 1.342 EHRs of non-operative and operatively managed adult patients. Data was split randomly into training and test set with the ratio 70-30. | Highest F1-scores achieved on test set across datasets: dataset 1: SciBERT: 0.387, dataset2: RandomForestClassifier with SMOTE oversampling: 0.383, dataset 3: RandomForestClassifier SMOTE: 0.378, dataset 4: RandomForestClassifier SMOTE: 0.383, dataset 5: RandomForestClassifier SMOTE: 0.383, dataset 6: RandomForestClassifier SMOTE: 0.388 |
| Comments: I do not think this paper is useful for my research in particular. Although it is a multi-class classification problem, the models were trained using very little data which clearly shows in the performance, yielding very low F1-scores. | | | | | | |
| 27 | Natural Language Processing for Automated Classification of Qualitative Data From Interviews of Patients With Cancer | Exploring the use of NLP methods for classifying free text from patient interviews in order to identify patient-reported symptoms. | BERT | The study comprises 3 interview datasets of transcripts from patients with liver cancer. Semi-structured interviews were conducted by interviewers and covered topic such as demographic background, disease background and symptoms. Quotations from the interviews were mapped to classes corresponding to the topics. Furthermore, detailed classification was done such as symptom and QoL impact-specific classifications. Four different NLP models were explored, these are TF-IDF, GloVe, RNN and BERT. Five BERT models were evaluated, each with a different number of transformer blocks (4, 6, 8, 10 or 12) | Data from three interview datasets, HCC with size 25, BTC with size 23 and GC with size 24. | The BERT model generally outperformed all other NLP models at both paragraph and sentence level. Mean ROC AUC on HCC dataset: 0.94 in predicting the classification of “symptom”, “QoL impact” and “other”. |
| Comments: This paper shows BERT models outperforming other NLP models on a classification task in the medical domain. As I decided to record F1-scores among papers and this paper does not show F1-scores it is hard to determine exactly how well the BERT models perform in comparison with other classification task. Furthermore, as they are not shown, it is hard to compare the performance of the other NLP models to the performance of BERT. | | | | | | |
| 29 | Deep Learning–based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports | Training a model to classify free-text oncology reports for tumour response category (TRC) | BERT | The data consists of examinations of all body regions from cancer research centres. Two NLP model types were trained. Type 1 applied a deep NLP algorithm based on BERT pretrained on the German vocabulary, which was finetuned on structured oncology reports (SOR). Type 2 served as NLP baseline and comprised Linear SVC, k-nearest neighbours and multinomial naïve bayes. A 5-fold cross-validation was performed on the SOR training set. Next to the NLP models The performance of BERT was compared to that of seven human annotators. | The final study sample included oncology reports from 10.455 patients. There are datasets with structured oncology reports and free-text oncology reports. | For the sake of this literature review we will consider the performance on the free-text oncology reports. Radiologists: ***0.79***, Medical students: 0.73, RT students: 0.64, BERT: **0.70**, Linear SVC: 0.63, K-nearest neighbors: 0.46, Multinomial naïve bayes: 0.56 |
| Comments: This paper shows that BERT is able to achieve comparable performance to skilled human annotators on a classification task in the medical domain. It also shows BERT greatly outperforming other NLP models on the same task. | | | | | | |
| 33 | Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing | Classifying social determinants of health from electronic health records by applying neural networks | BERT | Data stems from the MIMIC-III dataset, especially the “Social Work” report types. An ontology-based annotation scheme was made to label sentences in reports with 8 categories of social context. These serve as the classes for classification. Standard text preprocessing was applied to each sentence. Three deep neural network architectures are used, namely CNN, LSTM and BERT. The BERT model used the BERTbase preprocessing and pretrained transformer layers. A 256-node hidden layer was added to the BERT pooled output. This hidden layer was passed to an output layer with eight nodes. All models were evaluated over 10-fold stratified cross-validation. Two baseline methods are implemented as well, these being L2-regularized logistic regression and random forest classifiers. | The dataset used is MIMIC-III v1.4 which comprises over 58.000 hospital admissions. | Macro F1-scores: CNN: 0.55, LSTM: 0.555, BERT: **0.642** |
| Comments: This paper shows BERT outperforming CNN and LSTM models on the task of social determinants classification. The addition of the hidden layer could be a good idea for my research as well, even though it is not shown what the increase in performance is with compared to without this hidden layer. | | | | | | |
| 35 | Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach | Automatically extract staging-related information from CT reports to support accurate clinical staging of lung cancer. | BERT | Firstly, 22 questions about the diagnosis and staging of lung cancer were summarized. Subsequently, 14 types of entities and 4 types of relations were defined to represent the related information in the CT reports. A novel relation classification approach was proposed to determine relations between entities. Here, BERT is finetuned to fit the relation classification task. A new relation classification method is created using the relation sign constraint (RSC) to classify the relations between entities. Using this method, entity-entity relation constraints can be incorporated into a model to improve prediction performance. Next to BERT, an Attention-Bidirectional-LSTM model is also evaluated. BERT and this model are seen as baselines, while the RSC models are the novel ones proposed in this paper. | 392 chest CT reports. The data set was randomly divided into a training, validation and test set with the ratio 70-10-20. | Macro F1-scores on relation classification task: Attention-Bi-LSTM: 0.9574, BERT: 0.9631, Attention-Bi-LSTM-RSC: 0.962, BERT-RSC: **0.9713** |
| Comments: This paper shows the possible improvement in performance when adapting a relation sign constraint classification method. Although the dataset used was relatively very small the models performed reasonably well. It should be noted this is namely a relation extraction task which is very different from my task of lifestyle classification. It is unknown how well the application of relation sign constraint would perform there. | | | | | | |
| 36 | Predicting Glaucoma Progression Requiring Surgery Using Clinical Free-Text Notes and Transfer Learning With Transformers | Predicting glaucoma progression requiring surgery using ophthalmological clinical notes from electronic health records. | BERTbase, BioBERT v1.1 + PubMed, RoBERTa (base), DistilBERTbase | Clinical notes were preprocessed to remove stopwords and all letters were set to be lowercase. For the BERT models, the pretraining head of the model was removed briefly and replaced with a randomly initialized linear classifier that was passed through a softmax function to obtain the probability of progression to glaucoma surgery. All model parameters were finetuned on the training dataset. Evaluation is done on the test set. The output of each model is a probability score that must be converted to binary predictions based on a selected probability threshold. This threshold was tuned to maximize the F1-score. The performance of the BERT models was compared to that of a human glaucoma specialist. | Dataset includes three clinical progress notes per patient. The full dataset was split into a training, validation and test set with the ratio 8-1-1. The training set contains 3.612 patients, the validation set contains 400 patients and the test set contains 500 patients. | F1-scores on test set for predicting glaucoma progression: BERTbase: **0.45**, BioBERT v1.1 + PubMed: 0.42, RoBERTa (base): **0.45**, DistilBERTbase: 0.43, glaucoma specialist: 0.29 |
| Comments: This paper includes the conversion of an output probability to a class, which could be part of our new model, as we might implement confidence scores such as giving a score of 80% when the model is not entirely sure the patient in question is a smoker. | | | | | | |
| 61 | The natural language processing of radiology requests and reports of chest imaging: Comparing five transformer models’ multilabel classification and a proof-of-concept study | Comparing the performance of multiple (Dutch and English) BERT models on the task of multi-label classification on radiologist-annotated radiology requests. | BERTje, RobBERT, MultilingualBERT, ClinicalBERT, BERTbase, | In the pathology request dataset, each text item has multiple binary labels, making this a multi-label classification task. About 5% of the dataset was annotated manually. This annotated data was split into training and test sets with the ratio 90-10. The remaining unannotated data was sent through the best-performing request and report models to receive classification predictions. An LSTM model is used as baseline. For each of the five model types, specified in the previous column, three trained models were stored and used for evaluation. These models were the models with the best accuracy during training, the trained model after 2.000 iterations and the trained model after 4.000 iterations. | Data consists of 43.129 chest imaging requests and reports, all in Dutch. 2.256 of these were manually annotated and the remaining 40.873 were fed to the best-performing model to do classification on. | Among the five transformer models and one LSTM basline, the RobBERT model surpassed the others and was used for the multi-label classification of 40.873 radiology requests and reports. The RobBERT model that was trained for 4.000 iterations was chosen as best model based on its superior AUC in comparison with the other models. |
| Comments: This paper is rather interesting in the context of my research as it shows a Dutch model (RobBERT) outperforming English models on a multi-label classification task on clinical texts. Even though the Dutch model was not trained on medical texts it still outperformed ClinicalBERT on this task on particular. It could be the case that in my research clinical models like ClinicalBERT and BioALBERT are also worse than general Dutch models like RobBERT and BERTje. This is the reason I am comparing both English and Dutch clinical and general BERT models to my model. | | | | | | |
| 64 | Extracting Medication Changes in Clinical Narratives using Pre-trained Language Models | Exploring the automatic extraction of medication change information from free text clinical notes | BioBERT, ClinicalBERT, BlueBERT, PubMedBERT | This work uses the Contextualized Medication Event Dataset (CMED) to explore the automatic characterization of medication changes in patient records. Each mention of medication is assigned one or more event labels. These labels indicate whether there was a change in medication or not or whether a change cannot be determined. Furthermore, the label for change in medication (disposition) has five attributes, these being the action, actor, presence of negation, temporality and certainty. The CMED includes a training and test set with ratio 80-20. The task at hand can be conceptualized as two subtasks, these being medication mention extraction to determine when medication gets mentioned and event and attribute classification where mentions are labelled contextually. In order to tackle the problem of attribute classification five different architecture were tested, these being MedSingleTask, which has models for each event and attribute, MedMultiTask, which has one model for event and attribute predictions, MedSpan which can assign different labels to medications within the same sentence, MedIdentifiers which uses encoded input of medication and MedIDTypes which also has encoded input but uses typed markers. | The CMED dataset includes a training set with 400 notes and 7.230 medication mentions and a test set with 100 notes and 1.783 medication mentions. | Macro F1-score performance on event and attribute classification: MedSingleTask: 0.618, MedMultiTask: 0.651, MedSpan: 0.71, MedIdentifiers: **0.744**, MedIDTyped: 0.723 |
| Comments: This paper shows an approach for complex multi-label classification with multiple layers. Especially the negation part of the annotation is interesting for my research, as my model needs to be able to discern between negated and non-negated words. Annotated this negation however is a timely process and I need to consider if that is worth the time. | | | | | | |
| 66 | Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models | Presenting an automated method of information extraction from German unstructured clinical routine data from the cardiology domain. | GermanBERTbase, BERTfine, BERTscratch | For preprocessing, every discharge letter in de main corpus was converted to a raw text file. All patient personal data was deidentified as well. BERTbase and BERTscratch were pretrained on a single text file in which all discharge letters were concatenated. The corpus was tokenized using whitespace separation. Cardiologists added disease labels to the texts. Two baselines were used, these being conditional random field (CRF) and an RNN approach based on long-short term memory networks (LSTM). The standard BERT pretraining tasks are executed. After pretraining, the BERT models are finetuned as a supervised downstream task on annotated training data. Three different BERT models were evaluated: BERTbase which was based on a publicly available German BERT model trained on German Wikipedia and news texts, BERTfine which is based on BERTbase but finetuned using the entire corpus of German cardiology discharge letters and BERTscratch which was pretrained with the discharge letters from scratch. Results were measured on the CardioAnno corpus using 4-fold cross validation. | The main corpus contains 200.000 German discharge letters, around 2 GB of text | Average F1-score on CardioAnno corpus classifying: CRF: 0.83, LSTM: 0.80, BERTbase: **0.86**, BERTfine: **0.86**, BERTscratch: 0.83 |
| Comments: This paper shows that finetuning a model on a large set of data that was trained on another large set of data can outperform pretraining a BERT model from scratch. This tempts me to test this with my data as well, first build a finetuned model of another BERT model then compare it with my pretrained model. | | | | | | |
| 68 | Using deep learning-based natural language processing to identify reasons for statin nonuse in patients with atherosclerotic cardiovascular disease | Modifying ClinicalBERT to classify statin nonuse and reasons for statin nonuse using unstructured EHRs. | ClinicalBERT | The dataset consists if patients diagnosed with atherosclerotic cardiovascular disease (ASCVD). The dataset includes ICD-9 and 10 codes of various ASCVD disease variants. Clinical data was captured of the patients at the time of their diagnosis, such as age, gender, race/ethnicity, history of smoking, number of hospitalizations in the year prior to diagnosis and comorbidities. The primary outcome was statin nonuse based on structured and unstructured EHR data. The secondary outcome was classifying the reason for statin nonuse. The ClinicalBERT model was finetuned to classify patients as statin use versus nonuse. 10-fold cross-validation was applied to validate the model. | 56.530 patients with documented ASCVD who had at least two separate visits to the healthcare system. | F1-score on ASCVD dataset for classifying statin use: Binary classifier proposed in this paper: **0.85**, Random Forest baseline: 0.62 |
| Comments: This paper is rather vague, as it compares a BERT model solely to a Random Forest model on a classification task. Furthermore, it appears not many BERT alterations were made and this paper can be seen as a finetuning exploration task, but the performance is not compared to a not-finetuned model making it hard to discern what the effect of the finetuning was. For this reason, I will not consider this model any further. | | | | | | |
| 71 | Exploring Language Markers of Mental Health in Psychiatric Stories | Predicting whether patients are diagnosed with mental disorders and if so, the specific mental disorder type. | RobBERT | This exploratory study compares the classification performances of different NLP techniques and looks at which language cues could predict if a person has a mental disorder and if so, which disorder.The dataset for this study contains stories about mental illness of people who have or had psychiatric issues or were in contact with people with psychiatric issues. Diagnoses were assigned by multiple doctors and based on other material than the interviews. There were two tests in total. The first test consisted of classifying between mental disorder and no mental disorder. The second test consisted of classifying the specific disorder. RobBERT was evaluated on these tasks, as well as LIWC (Language Inquiry and Word Count, a tool often used for comparing mental disorders), spaCy (a software library often used for NLP) and fastText (a software library used for learning word embeddings). | The dataset consists of 108 interviews with 11 different diagnostic labels. The transcripts of the interviews are between 6.782 and 9.531 words long. A training test split of 80-20 applied. | Accuracy on classifying mental disorder vs no mental disorder: LIWC-Decision Tree: 0.857, LIWC-Random Forest: **0.952**, LIWC-SVM: 0.857, spaCY-Decision Tree: 0.81, spaCy-Random Forest: 0.762, spaCy-SVM: 0.714, fastText: 0.643, RobBERT: 0.607  Accuracy on multi-classifying mental disorder: LIWC-Decision Tree: 0.286, LIWC-Random Forest: 0.214, LIWC-SVM: 0.286, spaCY-Decision Tree: 0.143, spaCy-Random Forest: **0.429**, spaCy-SVM: 0.357, fastText: 0.286, RobBERT: 0.2 |
| Comments: This paper is very unique, as it shows a BERT model getting severely outperformed by older NLP methods. I believe this may have occurred because of the extremely small dataset size compared to other studies. Where other studies often use thousands to millions of texts, the database here is only 108 interviews big. This is a solid indicator that BERT models perform well on large datasets but can be outperformed massively on tasks with smaller datasets. As the dataset used in my research is very large I do not expect the same outcomes, but it is interesting to see what the performance is when I use the best-performing model (Random Forest) as a baseline. | | | | | | |
| 76 | Predicting COVID-19 Symptoms From Free Text in Medical Records Using Artificial Intelligence: Feasibility Study | Developing a prediction model to convert Dutch free text information on COVID-19-related symptoms from out of hours care electronic medical records into usable symptom-based data that can be analysed at large scale. | BERTje | This study is part of the project ID-CoV which aims to develop procedures for data identification, harmonization and linkage to develop robust methodologies to build a risk prediction tool for the identification of individuals at higher risk for severe COVID-19 outcomes. As input data, the iCAREdata database was used. The input data covers a broad range of different types of physicians of out of hours (OOH) care practices. The dataset contains records, where each record corresponds to one contact moment/consultation. For each record, 5 fields were extracted from which the “DiagnTekst” (diagnosis text) and “DiagnCod” (diagnosis code) are used as control records for validation. Two approaches for a multi-class classification model were considered, which were a classical machine learning-based text categorization approach and a deep neural network learning approach that is finetuned for classification. Preprocessing was done by removing empty entries that did not contain any information in the “objective” and “subjective” text fields. At first, a classic machine learning method was used, this being the Stochastic Gradient Descent classifier. Further experiments were then conducted with BERTje, a Dutch version of BERT. BERTje’s pretraining process was enriched with data from the iCAREdata database. The data from the dataset was split into training and test sets with the ratio 80-20. | Usage of iCAREdata dataset, containing roughly 779.000 records. Of these records 3.957 entries were randomly chosen to be annotated for the classification task. | Weighted F1-scores on test set: SGD classifier: 0.59, BERTje: **0.7** |
| Comments: This paper shows BERTje outperforming a standard machine learning approach on a classification task. It would be interesting to see whether other Dutch models like RobBERT would in turn outperform BERTje on the same task. It does show however that with a dataset size of only a few thousand of annotated data a BERT model is able to outperform the previous state-of-the-art. | | | | | | |
| 82 | Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing | Extracting the frequency of seizures from unstructured text in clinical notes | BioClinicalBERT, RoBERTa, BERT | For this study, progress notes from patients with epilepsy were retrieved. The problem-at-hand is a question-answering task where a model attempts to answer a series of questions given a context. Three questions were asked, these being: “has the patient had recent seizures?” (classification), “how often does the patient have seizures?” and “when was the patient’s most recent seizure?” (both text extraction). A sample from the full dataset was annotated by experts. This annotated dataset was split to training and test set with the ratio 70-30. On the test set, BERT serves as a baseline. BioClinicalBERT and RoBERTa were selected based on their improved performance on clinical benchmarking tasks. The finetuning pipeline uses a single pretrained model and performs unsupervised domain adaptation with MLM. Then, the models were finetuned on publicly available English datasets, where BoolQ3L was selected for the classification task. Performance was evaluated relative to ground truth annotations. | Initially 79.000 progress notes of patients with epilepsy. 1.000 of these notes were annotated by experts, which is what the finetuning classification task makes use of. | On the classification task, median F1-scores: BioClinicalBERT: **0.84**, RoBERTa: 0.745, BERT: 0.825 |
| Comments: This paper shows a comparison of multiple BERT models on classification in the biomedical domain. The big difference between the performance of BioClinicalBERT and RoBERTa is interesting to note. It is also interesting to note that BioClinicalBERT achieved a relatively high F1-score while only being finetuned on 700 records for the task, making the case for that model to be a strong candidate if I were to evaluate my model on an annotated subset of my data of around the same size or smaller. | | | | | | |
| 85 | Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods | Comparing the performance of several NLP models on the task of extracting diabetic retinopathy (DR)-related concepts from clinical narratives. | BERT, RoBERTa | A dataset of patients diagnosed with DR was used and annotated by a collaboration of experts. The goal is to extract DR concepts that can potentially help lesion detecting from medical images. Five different categories are defined. For the classification task, a standard two-stage NLP pipeline was adopted, which includes a clinical concept extraction module to detect DR-related attributes and a relation extraction module to link attributes to corresponding concepts. BERT and RoBERTa were evaluated on the tasks and compared to an LSTM model with classification layer, which functioned as baseline. The relation extraction pipeline consists of two steps including identifying pairs of concepts that might be related and classifying the relation categories using machine learning classifiers. Most relations occur in the same sentence, so heuristic rules are applied there to ensure only two concepts from the same sentence are considered to be related. For the BERT models, the base settings were examined, as well as the BERTmimic model trained on the clinical MIMIC-III database. NegEx was adopted to handle negations. In this literature review, I focus on the relation extraction task as it is most similar to my task. | A total of 536 fundus image reports from patients diagnosed with DR. From these reports, a total of 4.782 DR-related concepts are annotated. Around 73% of the data is used as training set and the remaining 27% as test set. | F1-scores on test set for relation extraction task: BERT: **0.9316**, RoBERTa: 0.9291, BERT trained with MIMIC-III: 0.9254, RoBERTa trained with MIMIC-III: 0.9304 |
| Comments: This paper shows BERT outperforming RoBERTa on a relation extraction/classification task, albeit not by a significant margin. The application of NegEx is an interesting concept and I believe it could be beneficial for my research to apply a Dutch version of NegEx in my research in order to boost performance. | | | | | | |
| 87 | Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study | Introducing a system for extracting family history information from clinical text. | BERT | This paper is a submission to the 2019 n2c2/OHNLP challenge on family history information extraction. The challenge includes two subtasks; subtask 1 includes recognizing family members with the side of the family, living status and observations in family history all from clinical text. For subtask 2 it is needed to extract the relations between family members, observations and living status and the task also includes negation classification. In this literature review, I focus on subtask 2 as it is most similar to my research task. The model used consists of two parts. The first part is a representation module, which represents input text using BERT and CNN-BiLSTM. The second part is a biaffine attention model to predict label score vectors. Two model settings were tested, these being a pipeline model that tackles unlabelled arc prediction and arc label prediction separately and a joint model that tackles these tasks concurrently. Furthermore, these models were compared to a BERT-BiLSTM model which uses BERT instead of CNN. | The training set included 149 documents. The development set includes 14 documents and the test set has 117 documents. | F1-scores on relation extraction task:  Pipeline model: 0.6851, Joint model: **0.7048**, BERT-2BiLSTM: 0.7009 |
| Comments: This paper gives an overview of an approach which was focussed on a particular NLP task. Due to its small training and test size I do not think the methods and performance of this paper is translatable to the context of my research. | | | | | | |

# BERT table

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Name paper** | | | **Type of task** | | **Name of (BERT) model(s) used** | **Methods** | | | **Type and amount of training data** | **Results** | |
| 1 | DocBERT: BERT for Document Classification | | | Finetuning BERT for document classification. | | BERTbase and BERTlarge | End-to-end optimization. Minimization of cross-entropy and binary cross-entropy loss. Finetuning on classification task, adding an additional output layer that signifies class label. | | | Four datasets, free text data, 80% training, 10% validation, 10% test data | F1-score on Reuters dataset test set: BERTbase: 0.89, BERTlarge **0.907**, LSTM 0.87, SVM 0.861, HAN 0.852 | |
| Comment: This was the first paper that demonstrated success in finetuning BERT for document classification and shows clearly how this can be achieved. The new structure proposed here can be useful if it shows up in newer BERT classification papers as well. | | | | | | | | | | | | |
| 9 | | Highly accurate classification of chest radiographic  reports using a deep learning natural language model  pre-trained on 3.8 million text reports | Classifying free text chest radiograph reports performed in the intensive care unit using several BERT models. | | BERTbase, GER-BERT, MULTI-BERT, newly trained model FS-BERT and RAD-BERT | | | Text extraction, where 948.000 of the 4.790.000 records were removed based on the fact they did not contain chest radiograph reports. Creating a custom WordPiece vocabulary for unique words, training data was generated using a maximum sequence length of 128 and a masked words percentage of 15%. The BERT models were pretrained by predicting masked words and doing next sentence prediction. Lastly, a small subset of data were randomly extracted and annotated. Nine phenomena were annotated, including congestion, opacity and effusion. Finetuning was done by refining the models for multi-label classification. | 3.841.543 reports went into pretraining of FS-BERT and RAD-BERT. Furthermore, about 7.200 text reports were randomly extracted and manually annotated by multiple people. If a text was deemed incomplete it was removed, leaving 5.203 final reports. These reports were split into a training set of size 4.703 and a test set of size 500. The training set was used to finetune the BERT models. | | On the largest training size F1-scores: RAD-BERT: 0.9, GER-BERT 0.89, MULTI-BERT: **0.92**, FS-BERT 0.85 | |
| Comments: This paper shows a rather surprising outcome, namely the multilingual BERT (MULTI-BERT) outperforming the BERT models that were directly pretrained (RAD-BERT and FS-BERT). Especially the difference between MULTI-BERT and FS-BERT is striking, as FS-BERT was built and trained from scratch. This paper shows therefore that it is possible that a finetuned BERT model that is pretrained on completely different data is able to outperform a BERT model that is trained solely on the source data. Of course this could also be a byproduct of the human annotation that was done to create the training and test sets, were MULTI-BERT could be more suited towards the structure of the labels than FS-BERT. I will not consider FS-BERT as a model to use in my research as it is very specific and gets outperformed by finetuned general models. | | | | | | | | | | | | |
| 11 | | A Pre-trained Clinical Language Model for Acute  Kidney Injury | | Predicting acute kidney injury (AKI) from unstructured clinical notes using a new BERT model (AKI-BERT). | | BERT, BioBERT, ClinicalBERT and newly trained AKI-BERT | | Data preparation, only patients were considered hat had a creatinine measured in 72 hours following ICU admission. The notes during  the first 24 hours of ICU admission for the selected patients are  extracted from MIMIC-III dataset. These are used for pretraining AKI-BERT. The dataset is imbalanced, and there are several measures used with the goal of dealing with this imbalance. AKI-BERT is pretrained similarly to the way Clinical BioBERT is pretrained. The finetuning is done by further training AKI-BERT on a single sentence classification task. | | The AKI corpus contains 77.160 clinical notes, which are used for pretraining AKI-BERT. | The best performance was achieved on a US + Pooling BERT structure. F1-scores: AKI-BERT: **0.451**, Bio+ClinicalBERT: 0.444, ClinicalBERT: 0.404 | |
| Comments: This paper is rather short and does not elaborate much on the process of creating AKI-BERT and the reason why the performance is lower than 0.5. I would say an F1-score of 0.451 does not warrant a good classifier, but other models (BioBERT, Clinical BERT etc.) also performed really poorly on this task which could indicate a problem with the input data. As AKI-BERT is trained on this potentially poor dataset I will not consider it further. | | | | | | | | | | | | |
| 12 | | Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing | Multi-label free text allergy chance classification | | mBERT, XLM-RoBERTa, WangchanBERTa, new RoBERTa-based model AllergyRoBERTa | | | AllergyRoBERTa was trained using a byte-level byte-pair encoding tokenizer. The pre-trained models were augmented with two linear layers with 768 hidden notes and 36 output nodes. Furthermore, ULMFiT and a combination of Naïve Bayes and Support Vector Machine is used to compare performance with. An ensemble model was also evaluated which chooses a label via majority vote among the other models. | 79.912 drug allergy records. Each allergy entry contains both Thai and English words. This dataset is divided into training, validation and test sets with a ratio of 80-10-10. There are a total of 36 classes a text can belong to and a text can belong to more than one class. | | | Performance on test set (F1-score): NB-SVM: 0.9695, ULMFiT: 0.9793, mBERT: 0.9577, XLM-RoBERTa: 0.9821, WangchanBERTa: 0.9753, AllergyRoBERTa: 0.9796, Ensemble model: **0.9888** |
| Comments: This paper shows an interesting method for creating an ensemble method. Rather than combining structures it compiles output and puts it to a majority vote, which in this case resulted in the best performance. Another interesting aspect is the XLM-RoBERTa model outperforming the model that was specifically pre-trained using the data that was tested on. This is similar to what occurred in paper 9. Even more interesting, the abundance of Thai language did not seem to hamper the models in the slightest as no text was translated to English. It does appear to be the case that XLM-RoBERTa handled this aspect better because it was pre-trained on multilingual data. It does need to be noted that AllergyRoBERTa was only trained on 1.2 million tokens, while XLM-RoBERTa was trained on billions. Nevertheless, the fact XLM-RoBERTa outperforms the model specifically trained on the source data and does this while the source data contains heaps of non-English text could very well indicate that XLM-RoBERTa could perform relatively well on my lifestyle classification task as well. | | | | | | | | | | | | |
| 17 | | Multiple Sclerosis Severity Classification From Clinical Text | | Pretraining a BERT model on consult notes about Multiple Sclerosis (MS) named MS-BERT, then finetune it to predict Expanded Disability Status Scale (EDSS) scores. | | MS-BERT, BlueBERT | | Their new MS-BERT model uses BlueBERT as a starting point. They use a masked language modelling pretraining task over deidentified consult notes. Consult notes are divided into chunks of 512 tokens so longer notes can be processed. They created a custom classifier named MSBC to predict MS severity labels. MSBC first reads in a consult note, then tokenizes the text using the BERT vocabulary and then splits the tokens into chunks of 12. Semi-supervised labelling was applied using the Snorkel framework, which makes use of weak supervision. | | The dataset was compiled by a leading MS research hospital and contains approximately 70.000 consult notes. Of these, 16.000 were labelled manually. After deidentification the labelled set was separated into a training, validation and test set in a 50-20-30 ratio. | Micro F1-scores on test set: MSBC: **0.94177**, MSBC with only first 512 tokens: 0.90086, Rule-Based keyword search + MSBC: 0.92987 | |
| Although MS-BERT is most likely not useful for my lifestyle classification task, there is a lot to learn from its pretraining and finetuning parts. Splitting the notes to parts of 512 tokens is a strategy I will most likely also need to apply to my clinical notes, as a lot of them will exceed 512 tokens in length. The Seq2Vec encoder, which is a part of MSBC, could increase performance in my finetuning step as well, which would need to be tested. | | | | | | | | | | | | |
| 20 | | BERTje: A Dutch BERT Model | | Pretraining a BERT model on a large corpus of Dutch text | | Multilingual BERT, BERTje | | Because of the demonstrated ineffectiveness of the NSP task during pretraining, BERTje is trained with the SOP objective, which means the second sentence in each training example is either the next or the previous sentence. A different strategy is also applied to the MLM objective where instead of randomly they mask consecutive word pieces that belong to the same word. The final model is pretrained for 1 million iterations. | | For pretraining, several corpora of Dutch text were combined, such as books, Dutch news, multi-genre references, all articles of 4 Dutch news websites from 2015 to 2019 and the October 2019 Wikipedia dump. This totals to about 12 gigabytes and about 2.4 billion tokens. | Three models were finetuned for three different tasks. These models are multilingual BERTbase, BERTje at 850.000 iterations and BERTje at 1 million iterations. For the semantic role labelling task in particular, the F1 scores on the test set of the CoNLL-2002 task were: multilingual BERTbase: 0.804, BERTje850k: 0.852, BERTje: **0.853** | |
| Comments: This model is very useful for my research. I can use it directly to finetune it on my classification task and record performance that way in order to compare that performance with the performance of my pretrained model. The fact that it is trained using Dutch source data is a really big plus, as my source data is also largely in Dutch. A possible hinderance could be that there presumably are not a lot of medical terms in the source data which would occur in my sources. | | | | | | | | | | | | |
| 22 | | RobBERT: a Dutch RoBERTa-based Language Model | | Training a robustly optimized BERT model for the Dutch language. | | BERTje, BERT-NL, RobBERT, mBERT | | RobBERT is pretrained using the RoBERTa training regime. Two different versions were trained, one where only the pretraining corpus was replaced with a Dutch corpus, named RobBERT v1 and one where both the corpus and the tokenizer with replaced with Dutch versions, named RobBERT v2. The vocabulary of the Dutch tokenizer was constructed using the Dutch section of the OSCAR corpus. Only the MLM task is used during pretraining, just like in RoBERTa. Evaluation is done by testing the new model on sentiment analysis and pronoun prediction, among other tasks. For the sentiment analysis task RobBERT was finetuned on Dutch book reviews. | | RobBERT was trained on a 39GB Dutch corpus of 126 million lines of text. The corpus is much larger than BERTje (12GB) and BERT-NL (2.2GB). | F1-scores on sentiment analysis task:  BERT-NL: 0.84, RobBERT v1: 0.94422, RobBERT v2: **0.95144**  Accuracy of RobBERT v2 was higher than BERTje on sentiment analysis.  F1-scores on pronoun prediction: mBERT: 0.98033, BERTje: 0.98014, RobBERT v1: 0.98169, RobBERT v2: **0.99121** | |
| Comments: Like BERTje, this model also seems very useful for my research and could even be more useful due to its larger training corpus and higher achieved performance on the sentiment analysis and pronoun prediction tasks. As a possible negative point again there might not be as many medical terms or the input dataset might not be focussed on medical terms enough to perform better than BERT models that are trained on medical corpora. | | | | | | | | | | | | |
| 23 | | RobBERTje: A Distilled Dutch BERT Model | | Training a distilled version of the state-of-the-art RobBERT model for Dutch language | | RobBERT, BERTje, RobBERTje | | Distilling the model gets rid of a lot of the redundant heads in the large BERT architecture. To perform distillation, they distil several smaller models from the Dutch RobBERT model, which are the RoBERTa architecture and the OSCAR corpus. Optimal hyperparameters found in the Bort research are tested. It is also tested if it matters whether the training corpus is shuffled, what the influence of the length of the training sequences is and what distillation architecture hyperparametrization works best for the distilled model. The model is evaluated on the same tasks as in the RobBERT paper, being sentiment analysis, co-reference resolution and named entity recognition. | | The model is pretrained using 1 GB of data, greatly reducing upon the 39GB of RobBERT. For the sentiment analysis, the Dutch Book Reviews Dataset (DBRD) is used. This is a binary classification dataset which contains 22.000 book reviews. | Accuracy on sentiment analysis: RobBERT v2: **0.944**, BERTje: 0.93, best RobBERTje version: 0.929 | |
| Comments: This is an interesting paper in that it does not necessarily propose a method that improves upon the state-of-the-art in terms of performance, as on almost all of the tasks the performance was worse than the state-of-the-art. What this paper does show is that a similar performance can be achieved with an extremely significantly smaller model size. Which means this model is way more usable in real-world applications as it is quicker and smaller. For my research, I believe the resources I have access to will largely decide whether I will evaluate my new model against a big model like RobBERT or distilled versions like RobBERTje. Nevertheless, I deem the chances high that this model will be used in evaluation. | | | | | | | | | | | | |
| 24 | | BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | | Introducing a new language representation model called BERT (Bidirectional Encoder Representations from Transformers) | | BERTbase, BERTlarge | | The BERT framework includes two steps: pretraining and finetuning. During pretraining the model is trained on unlabelled data over two different pretraining tasks. For finetuning the parameters are finetuned using labelled data from downstream tasks. BERTbase and BERTlarge are introduced, where base includes around 110 million parameters and large includes 340 million. BERT can represent both a single sentence and a pair of sentences in one token sequence. Pretraining is done by masked-language modelling and next sentence prediction. For finetuning, the appropriate inputs and outputs are swapped out, such as adding output layers for classification. Within this literature review we focus on the GLUE task for results as it is a classification task. This task adds an output layer with an amount of nodes that corresponds to the amount of labels to classify. | | For pretraining they use the BooksCorpus, which includes 800 million words, and the entirety of English Wikipedia (2.5 billion words). | On the GLUE task, average accuracies on all subtasks were: Pre-OpenAI SOTA: 0.74, BiLSTM+ELMo+Attn: 0.71, OpenAI GPT: 0.751, BERTbase: 0.796, BERTlarge: **0.821** | |
| Comments: As the original source paper for BERT, this paper is very helpful in describing the intricacies that went into creating the first BERT model, as well as providing reasons for why certain choices were made. I will definitely use this model as one of my baselines. | | | | | | | | | | | | |
| 37 | | ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission | | Introducing a new BERT model pretrained using clinical notes. The model is then finetuned on the task of predicting hospital readmission. | | ClinicalBERT, BERT | | In ClinicalBERT, a token in a clinical note is represented as a sum of the token embedding, a learned segment embedding and a position embedding. The segment embedding identifies which sequence a token is associated with. The position embedding of a token is a learned set of parameters corresponding to the token’s position in the input sequence. ClinicalBERT uses the same pretraining tasks as the original BERT model. The model is finetuned on readmission probability prediction. Here, admissions where a patient is readmitted within 30 days are labelled 1 and all other patients are labelled 0. | | The MIMIC-III dataset is used, which includes 2.083.180 deidentified clinical notes. The final cohort of the 30-day hospital readmission prediction task includes 34.560 patients with 2.963 positive and 42.358 negative readmission labels. For validation and testing, the cohort is split into five folds. Within each fold, 20% of the data is used for validation, 10% for testing and the rest for training. | AUROC on readmission prediction task:  ClinicalBERT: **0.714**, Bag-of-words: 0.684, Bi-LSTM: 0.694, BERT: 0.692 | |
| Comments: This paper is also very useful as a baseline as it is trained on clinical notes and my training data also comprises of clinical notes. It is interesting to explore whether this model outperforms BERT on Dutch clinical note classification as well. | | | | | | | | | | | | |
| 38 | | Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training | | Introducing Medical Vision Language Learner MedViLL with the goal of maximizing generalization performance for vision-language understanding and generation tasks. | | PixelBERT, BERT | | The proposed architecture MedViLL is a single BERT-based model that learns unified contextualized vision-language representation. A convolutional neural network is used to extract visual features from medical images. A BERT model is used to embed language features. These visual and language features are concatenated to construct the input sequence to the joint embedding component. The pretraining tasks include Masked Language Modelling and Image Report Matching. Several types of self-attention masks are explored to encourage the model to learn universal multimodal representations. Ultimately, four different models are implemented each using different types of self-attention masks: Bi, S2S, BAR and Bi&S2S. MedViLL is compared to these models as well as two baselines, these being “finetuning only” which has the same architecture as MedViLL but is not pretrained, and “CNN & Transformer”, which uses the CNN module for encoding images and a Transformer module for encoding reports. | | The MIMIC-CXR and Open-I datasets were used. The first dataset contains 377.110 chest X-ray images and corresponding free text reports. The Open-I dataset contains 3.851 reports and 7.466 chest X-ray images. For the MIMIC-CXR the official split of training, validation and testing is used. Open-I is used to test the generalization ability of the model, which is pretrained on the MIMIC-CXR dataset. | Average F1-scores for diagnosis classification task on MIMIC-CXR dataset: MedViLL: 0.839, Bi&S2S: 0.846, Bi: **0.852**, S2S: 0.846, Finetuning only: 0.807, CNN&Transformer: 0.491 | |
| Comments: I will most likely not use any of the ideas proposed in this paper. Although a lot of alterations to regular BERT are proposed, the model does not outperform the baselines on the task it was pretrained on. Furthermore, the model was trained solely on X-ray reports which does not bode well for performance on other clinical notes, especially as it gets outperformed on its own task. | | | | | | | | | | | | |
| 39 | | VisualCheXbert: Addressing the Discrepancy Between Radiology Report Labels and Image Labels | | Producing labels from radiology reports that have better agreement with radiologists labelling chest X-ray images | | BERTbase, VisualCheXbert | | The methods differ among the six tasks in this paper. We focus on the last task, as it focusses directly on mapping free text radiology reports to X-ray image labels. Within this task, a DenseNet model is trained to detect medical conditions from the images, as manual annotation is costly. The DenseNet model outputs probability for each of the 14 conditions. These probabilities are used as ground truth to finetune a BERTbase model. One BERT model is trained on the MIMIC-CXR dataset and another on the CheXpert training dataset. BERT’s output is fed to linear heads, one head for each medical condition. After training the BERT model, the output probabilities are mapped to positive or negative labels for each condition either by converting the outputs to binary labels and maximizing the sum of sensitivity and specificity minus one (BERT+Thresholding) or by training a logistic regression model for mapping (BERT+LogReg). Performance of these methods is compared. | | Usage of MIMIC-CXR dataset, which consists of 377.110 chest X-rays and corresponding radiology reports. Furthermore, the CheXpert dataset is used which contains a separate set of 700 chest X-ray validation and test studies labelled by radiologists. F | F1-scores for BERT models trained on MIMIC-CXR dataset: BERT+Thresholding with DenseNet Labels: 0.65, BERT+LogReg with DenseNet Labels (VisualCheXbert): **0.73**.  F1-scores for BERT models trained on CheXpert dataset: BERT+Thresholding with DenseNet Labels: 0.65, BERT+LogReg with DenseNet Labels: **0.72**. | |
| Comments: The methods of thresholding in this paper show promise if I decide to work with confidence probabilities for my model as well. For example, I could apply logistic regression to my output probabilities based on the findings of this paper in order to distinguish between smokers and non-smokers. I do not think I will use the VisualXbert model furtherly as it is focussed on X-ray reports and very likely is not very adaptable to the whole of Dutch clinical texts. | | | | | | | | | | | | |
| 40 | | RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification | | Pretraining transformers using contrastive learnng before the end-to-end finetuning to improve multi-class classification of radiology reports. | | BERT, BlueBERT, RadBERT-CL | | The model classifies radiology reports into different disease observations. For every possible class the model outputs either: blank, positive, negative or uncertain. The data for the model is preprocessed on sentence and document level. To do this an Info-Preservation module is defined, which identifies and preserves facts during augmentation so that valuable data does not get lost. Within this model, to capture the presence of negation of any concept, a dictionary of 30 negation indicator keywords is created manually. These keywords include words like “not” and “without”. Sentence augmentation is done by splitting reports to sentences and them applying random word/phrase dropping. Document-level augmentation is done by removing extra spaces, newlines and unwanted tokens.  For pretraining the RadBERT-CL model, the BlueBERT architecture is used as an encoder. The pretraining is done using contrastive self-supervised learning. Three novel contrastive learning algorithms are proposed for RadBERT-CL, these are patient-based document level CL, disease-based sentence level CL and disease+factuality-based sentence level CL. The finetuning is done on pseudo-labels of a radiology report classification task. | | The MIMIC-CXR dataset is used, which consists of 377.110 chest X-ray images. The dataset is pseudo-labeled using an automatic labeller for the intended set of 14 observations. The dataset is divided into training and test sets with the ratio 80-20. | Weighted F1 scores on test set for finetuned RadBERT-CL variants, compared to CheXbert, a state-of-the-art radiology model:  CheXpert: 0.743, previous state-of-the-art CheXpert: 0.798, RadBERT-CL (patient-based document level CL): 0.799, RadBERT-CL (disease-based sentence level CL): 0.801, RadBERT-CL (disease+factuality-based sentence level CL): **0.804** | |
| Comments: This paper is interesting because its data augmentation phase takes negation into account. Negation is something that I will have to deal with in my model as well and the methods proposed for it here could be helpful in that regard. The application of contrastive learning is also something that I can consider for my model, and the clear examples and results achieved using three different contrastive learning approaches could be very useful if I were to test the effect of applying contrastive learning to my model. | | | | | | | | | | | | |
| 41 | | TinyBERT: Distilling BERT for Natural Language Understanding | | Proposing a novel transformer distillation method that can transfer part of the knowledge from BERT to a significantly smaller model named TinyBERT. | | BERTbase, BERTtiny, BERTsmall, BERT-PKD, DistilBERT, MobileBERTtiny, TinyBERT | | The aim of knowledge distillation (KD) is to transfer the knowledge of a large teacher network to a small student network. Within KD, the objective is to minimize the difference between these networks. Here, a novel distillation method is proposed wherein both the student and teacher networks are built using transformers. A predefined amount of layer is distilled from the teacher and mapped to the student. Attention based distillation is utilized, which can capture rich linguistic knowledge that can be transferred to the student. Furthermore, distillation is also performed on the embedding and prediction layers. Pretraining includes general and task-specific distillation from BERT. General distillation is done by using the BERT model without finetuning as teacher, while task-specific distillation takes the finetuned BERT as teacher. | | TinyBERT learns directly from the BERTbase model, which contains 109 million parameters. English Wikipedia is used as the text corpus. | Performance was measured on the GLUE tasks. Average accuracies on test set: BertBASE (teacher): ***0.795***, BERTtiny: 0.702, BERTsmall: 0.721, BERT-PKD: 0.726, DistilBERT: 0.719, MobileBERTtiny: **0.77**, TinyBERT: **0.77** | |
| Comments: This paper shows a comparison between multiple distilled versions of BERT and proposes a new model that, despite having significantly less layers, is not far off the original BERT’s performance. It performs similarly to MobileBERTtiny, which has 24 layers while TinyBERT only has 4. This model is useful for my research if the resources I have access to are limited, as it is way smaller and faster than regular BERT. | | | | | | | | | | | | |
| 42 | | SpanBERT: Improving Pre-training by Representing and Predicting Spans | | Introducing a new pretraining method for BERT models that is designed to better represent and predict spans of text. | | SpanBERT, BERT | | SpanBERT differs from BERT as it uses a different random process to mask spans of tokens rather than individual tokens. To do span masking, spans of text are sampled iteratively until a masking budget has been spent. At each iteration, a span length is sampled from a geometric distribution. Then, a starting point for the mask in the span is selected randomly. Instead of the next sentence prediction task of BERT, SpanBERT uses only MLM for pretraining. SpanBERT is evaluated on the SQuAD 1.1 and 2.0 benchmarks, which are about extractive question answering. The model is furthermore evaluated on five more datasets from several tasks. For SpanBERT, the model configuration of BERTlarge was applied. The main differences are SpanBERT using different masks at each epoch as opposed to sampling 10 different masks for each sequence, and removing all short-sequence strategies. SpanBERT is compared to the original BERT model, the authors’ reimplementation of BERT with improved data preprocessing and optimization and the authors’ reimplementation of BERT trained on full-length sequences without NSP. | | Training was done on the same data as BERTlarge was, this being BooksCorpus and English Wikipedia. | F1-scores on SQuAD 1.1 test set: Human performance: 0.912, Original BERT: 0.913, Authors’ BERT: 0.926, Authors’ BERT without NSP: 0.933, SpanBERT: **0.946**  F1-scores on SQuAD 2.0 test set:  Human performance: ***0.894***, Original BERT: 0.833, Authors’ BERT: 0.859, Authors’ BERT without NSP: 0.866, SpanBERT: **0.887** | |
| Comments: This paper shows an improvement upon BERT by changing its pretraining procedure. It would be very interesting to see whether the focus on spans increases performance on my dataset as well. On the SQuAD 1.1 benchmark the model even outperforms humans, which shows a lot of promise. Because it is a general model I think this model is fit for being a model I compare my model’s performance with. | | | | | | | | | | | | |
| 44 | | HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language | | Proposing a way to improve the performance of BERT on the detection of abusive language. | | BERT, HurtBERT-Enc, HurtBERT-Emb | | The paper proposes two models. Both models start with the BERT layer, this BERT layer connects to a dense layer. Regarding the features from the hate speech lexicon used (HurtLex), the models differ from each other in the way they extract features. The first model makes use of encoding (HurtBERT-Enc), while the second model uses embedding with an LSTM (HurtBERT-Emb). A big difference is that embedding is on word-level while encoding is on comment (document) level. For both models, the dense layer from the BERT output and the dense layer from the HurtLex are concatenated and output into a prediction layer. The HurtBERT models are compared to the °bert-uncased” model from tensorflow-hub. The models are evaluated on six datasets, which are all in English and mostly stem from research on tweets. HurtBERT is also trained using these datasets and performance is measured while using different datasets for training and testing. | | The data the model is trained on depends on the size of the respective datasets, which are:  Waseem: 16.914 tweets, Davidson: 24.783 tweets, Founta: 80.000 tweets, HatEval: 12.000 tweets, OLID: unknown quantity from paper, AbuseEval: unknown quantity from paper. | Performance was recorded using different datasets for training and testing, which makes it hard to show full results. In summary, HurtBert-Enc and HurtBert-Emb outperformed BERT with every training dataset evaluated on every test dataset. HurtBERT-Emb has overall the best performance. | |
| Comments: I do not believe this model is useful for my research as its topic is far removed from mine and it does not provide as interesting alterations to regular BERT as other papers. I will therefore not consider it further in the evaluation and creation of my model. | | | | | | | | | | | | |
| 45 | | ALBERT: a lite BERT for self-supervised learning of language representations | | Presenting a new model that lowers the GPU/TPU memory load of BERT and results in a much more scalable model | | BERTbase, BERTlarge, ALBERTbase, ALBERTlarge, ALBERTxlarge, ALBERTxxlarge | | ALBERT uses a pretraining loss based on predicting the ordering of two consecutive sentences of text. This differs from the BERT pretraining approach, that includes predicting whether the second segment in a pair has been swapped with a segment from another document. Several other design choices have been made to improve upon BERT. One such choice is the application of factorized embedding parameterization, which decomposes the embeddings into two smaller matrices, reducing the embedding parameters in size compared to WordPiece embeddings used in BERT. Furthermore, cross-layer parameter sharing is proposed as a way to improve parameter efficiency, increasing smoothness of transitions between layers. Finally, next sentence prediction is replaced with sentence-order prediction which is a more complex task. In total, four new models are created, which differ in the amount of layers and input parameters. The performance of these models is compared to BERT on several state-of-the-art NLU tasks, such as the GLUE benchmark. | | ALBERT is trained on the same data as BERT in order to make comparison easier. This data consists of the English Wikipedia and BookCorpus datasets. | For singular models, ALBERT trained with 1.5 million steps outperforms BERTlarge, XLNetlarge, RoBERTa-large and ALBERT (at 1 million steps) in every task. For ensemble models, the ALBERT ensemble consisting of models of various step sizes outperforms ALICE, MT-DNN, XLNet, RoBERTa and Adv-RoBERTa on 7 of the 9 tasks and has the highest average score by 0.6% over Adv-RoBERTa. | |
| Comments: The proposed changes to BERT’s pretraining structure plus the fact this model outperforms the state-of-the-art on the GLUE benchmark shows a lot of promise. The fact that it is trained on the same data as BERT and is able to outperform it with less parameters clearly shows the significance of the pretraining alterations. I will consider the same alterations for my model and will finetune ALBERT on my data to compare performance to my model. | | | | | | | | | | | | |
| 46 | | BioBERT: a pre-trained biomedical language representation model for biomedical text mining | | Adapting BERT to perform better on biomedical corpora | | BERT, BioBERT v1.0, BioBERT v1.1 | | BioBERT is initialized using BERT’s weights then pretrained on biomedical domain corpora (PubMed abstracts and PMC full-text articles). This makes the model domain-specific for medical purposes. BioBERT is finetuned on three representative biomedical text mining tasks: named entity recognition, relation extraction and question answering. Two BioBERT models are proposed v1.0 and v1.1 which differ in the amount of pretraining steps (270 thousand in 1.0 compared to 1 million in 1.1) | | BioBERT is trained one four corpora: English Wikipedia (2.5 billion words), BooksCorpus (800 million words), PubMed Abstracts (4.5 billion words) and PMC Full-text articles (13.5 billion words). | In terms of F1-score BioBERT models outperformed BERT models on 5 of the 8 datasets of the biomedical named entity recognition task. | |
| Comments: Because this model was trained on biomedical corpora it is interesting for my research, and I will use it to compare my model’s performance to. In terms of alterations to the original BERT this paper is less interesting and I will not use it as a reference for that. | | | | | | | | | | | | |
| 47 | | BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension | | Applying a denoising autoencoder for pretraining sequence-to-sequence models in a new model named BART. | | BERTbase, BARTbase | | BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from. For pretraining the negative log likelihood of the original document is optimized. Documents are corrupted and then reconstructed. The addition of several transformations are tested, such as token masking like in BERT, token deletion, text infilling from SpanBERT, sentence permutation and document rotation. Finetuning for a sequence classification task is done by feeding the same input to the encoder and decoder and then feeding the final decoder token to a new multi-class linear classifier. BART is also finetuned on the tasks of token classification, sequence generation and machine translation. | | BART is pretrained on the same training data as RoBERTa, this being 160 GB of news, books, stories and web text. | F1-scores on SQuAD 1.1 task: BERT: 0.909, XLNet: 0.945, RoBERTa: **0.946**, BART: **0.946**  F1-scores on SQuAD 2.0 task: BERT: 0.818, UniLM: 0.834, XLNet: 0.888, RoBERTa: **0.894**, BART: 0.892 | |
| Comments: This model is very comparable to RoBERTa, but performs slightly worse most likely due to the addition of text generation capabilities. As text generation is not relevant for my research I will consider RoBERTa rather than BART. | | | | | | | | | | | | |
| 49 | | EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain | | Producing a BERT model named EntityBERT focused on the clinical domain by continued pretraining of a model with broad representation of biomedical terminology. | | BioBERT, PubMedBERT, BARTlarge, PubMedBERT, EntityBERT | | PubMedBERT has shown to outperform BioBERT, ClinicalBERT, RoBERTa and BARTlarge. In this study, PubMedBERT is furtherly pretrained on a clinical corpus and a novel entity-based masking strategy is used to infuse domain knowledge in the learning process. The pretraining is furthered using the MIMIC-III clinical dataset. This corpus is processed using the sentence detection, tokenization and temporal modules of Apache cTAKES. The entity-based masking replaces the MLM pretraining task from BERT. Within this strategy, all entities in the input sequence are marked with XML tags, which are all marked with unique IDs. Then, 40% of entities and 12% of random words are chosen respectively within each sequence block for corruption, following a 80-10-10 ratio for replacing with [MASK], keep it unchanged or replace it randomly. The NSP task was not used entirely to pretrain EntityBERT. The pretrained model is finetuned for three clinical tasks: DocTimeRel classification, temporal relation extraction and negation detection. | | EntityBERT is built upon the PubMedBERT model, which has been pretrained from scratch on 21 GB of PubMed text. The EntityBERT model is further pretrained on the MIMIC-III clinical dataset which consists of approximately 2 million deidentified notes. | Overall F1-scores on the TLINK label classification task: BioBERT: 0.625, BARTlarge: 0.628, PubMedBERT: 0.638, EntityBERT: **0.651** | |
| Comments: This paper shows a way in which a existing model can be improved by further pretraining it on other data. For my model, I could choose to further improve an existing model with my data instead of pretraining a model from scratch if the former yields better performance. The entity-based masking method is also something that I could consider for my model. | | | | | | | | | | | | |
| 50 | | A disease-specific language representation model for cerebrovascular disease research | | Exploring whether a BERT-based model pretrained on disease-related clinical information, StrokeBERT, can be more effective for cerebrovascular disease-relevant research. | | StrokeBERT, ClinicalBERT | | StrokeBERT is trained on data of cerebrovascular disease patients. All clinical notes were preprocessed by removing Chinese characters, special characters and multiple spaces. StrokeBERT was initialized with BioBERT’s weights. The size of StrokeBERT is the same as the BERTbase model and the same pretraining tasks were used, these being MLM and NSP. The BERT WordPiece tokenizer was used for tokenization. After training, the model can be finetuned by adding an additional output layer to create task-based models such as disease identification. StrokeBERT’s performance is measured in comparison to ClinicalBERT. Performance was measured on several tasks , where we focus on the classification of extracranial and intracranial artery stenosis from radiology reports. In this task, 80% of the 9.614 total reports were used to finetune the model and the other 20% was used for validation. | | Clinical notes and reports from the Chang Gung Research Database were used. In total 113.590 discharge notes, 105.743 radiology reports and 38.199 neurological reports. | StrokeBERT improved upon ClinicalBERT on both extracranial and intracranial artery stenosis identification, albeit not significantly. AUC: StrokeBERT: **0.968**, ClinicalBERT: 0.956 | |
| Comments: This paper is another relatively very specific BERT model which is not general enough in the medical domain and does not improve enough upon the baseline to be interesting enough to include in my research further. The alterations to regular BERT were also unremarkable. | | | | | | | | | | | | |
| 52 | | RoBERTa: A Robustly Optimized BERT Pretraining Approach | | Improving BERT by making different design choices on the aspects of hyperparameters and training size. | | BERTbase, BERTlarge, RoBERTa | | In the paper, the authors show that removing the NSP pretraining task matches or slightly improves downstream task performance for BERT. Secondly, as BERT relies on randomly masking and predicting tokens in a way which sees a lot of duplicate masks, also known as static masking, the authors wanted to see whether a more dynamic masking pattern would be beneficial. By generating the masking pattern every time a sequence is fed to the model dynamic masking can be achieved, which is more scalable to larger models and performs similarly to static masking BERT. Furthermore, by swapping the character-level byte-pair encoding from BERT with a larger byte-level vocabulary, adding 15 million and 20 million parameters for BERTbase and BERTlarge respectively. This adds the advantages of using a universal encoding scheme. Overall, RoBERTa is trained with dynamic masking, full sentences without NSP loss and larger byte-level byte-pair encoding. Furthermore, the effect of having more training data and having the data passed more often through the model is measured. RoBERTa is evaluated on the GLUE benchmark as well as on the SQuAD datasets. | | RoBERTa was trained on the BookCorpus plus English Wikipedia dataset, just like regular BERT. | For the single-task singular models, RoBERTa outperformed BERTlarge and XLNet (large) on every single task. For the ensemble models on the test sets, the following average scores were achieved: ALICE: 0.863, MT-DNN: 0.876, XLNet: 0.884 and RoBERTa: **0.885** | |
| Comments: RoBERTa is used in a lot of other papers that introduce BERT models and its proposed changes to BERT are shown to outperform the state-of-the-art at the time. For this reason, I will most likely consider techniques like dynamic masking and leaving out NSP and use this model to compare my model to in terms of performance on the finetuning task of multi-label lifestyle characteristic classification. | | | | | | | | | | | | |
| 54 | | UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus | | Introducing a contextual embedding model, UmIsBERT, that integrates domain knowledge during the pretraining process via a novel knowledge augmentation strategy. | | BERTbase, BioBERT, BioClinicalBERT, UmIsBERT | | In the UmlsBERT model the MLM procedure is updated to take the associations between the words specified in the UMLS Metathesaurus into account. The UMLS Metathesaurus is a compendium of many biomedical terminologies with associated information like synonyms and categorical groupings. For UmlsBERT, a new embedding matrix is introduced into the input embedding which captures the unique semantic meaning of the words in UMLS. Furthermore, the loss function of the MLM pretraining task is updated so it can handle multi-label classification. Instead of using a 1-hot vector for singular words a binary vector indicating the presence of all similar words are used. UmlsBERT is initialized with the pretrained BioClinicalBERT model then further pretrained with the updated MLM task on MIMIC-III notes. Finetuning is done by adding a single linear layer on top of UmlsBERT. UmlsBERT is trained for 1 million steps. UmlsBERT is compared to BERTbase, BioBERT and BioClinicalBERT. | | The MIMIC-III dataset is used to pretrain UmlsBERT, in particular it is trained on the “NO-TEEVENTS” table, which contains 2.083.180 rows of clinical notes and test reports. This table includes five datasets and performance is measured on the test set of each of these datasets. | Test accuracies on MedNLI dataset: BERTbase: 0.779, BioBERT: 0.822, BioClinicalBERT: 0.812, UmlsBERT: **0.83**  i2b2 2006: BERTbase: 0.935, BioBERT: 0.933, BioClinicalBERT: 0.934, UmlsBERT: **0.944**  i2b2 2010: BERTbase: 0.852, BioBERT: 0.873, BioClinicalBERT: 0.877, UmlsBERT: **0.886**  i2b2 2012: BERTbase: 0.765, BioBERT: 0.778, BioClinicalBERT: 0.789, UmlsBERT: **0.794**  i2b2 2014: BERTbase: **0.952**, BioBERT: 0.946, BioClinicalBERT: 0.943, UmlsBERT: 0.949 | |
| Comments: As this model is focussed on quite a unique and specific task, this being the UMLS Metathesaurus, and altered the BERT structure based on that I will not consider this model or these alterations for my model. I will also not use this model for evaluation. | | | | | | | | | | | | |
| 55 | | A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study | | Developing a BERT-based system to automatically extract detailed tumor site and histology information from free-text oncological pathology reports, | | ClinicalBERT, caBERT | | The aim of the study is to extract accurate tumour site and histology descriptions from complex free-text pathology reports, accommodate the terminology used to indicate the same pathology and provide accurate standardize tumour site and histology codes for use by downstream applications. The model was initialized using the weights from ClinicalBERT. The model is a Question&Answer model which appends a Q&A layer onto a pathology language model to answer pathology questions. This pathology model was trained in three stages, first learning how to answer general language questions, then answer technical biomedical science questions and finally answer questions from pathology reports. Each stage makes use of supervised learning, which requires ground truth labels. Finetuning was done on 16.782 reports of the SQuAD Q&A dataset, which were split into a training and test set with the ratio 80-20. ClinicalBERT was finetuned on the same task and serves as a baseline. | | caBERT was trained on electronic pathology reports of solid tumors produced by Moffitt. In total 275.604 reports. | F1-scores on answering pathology-related questions from the Moffitt dataset: ClinicalBERT: 0.8485, caBERT: **0.8776** | |
| Comments: This model presents a way to alter BERT to work on a Q&A task, especially in the pathology/oncology field. It shows that using the weights of ClinicalBERT and further training a model on a specific task could lead in improvement on that specific task. As this model is again quite specific to one particular field (cancer research) I will not consider the model further in my research. I might however test the process of initializing my model using weights of another model, as this seems to boost performance in multiple studies. | | | | | | | | | | | | |
| 59 | | Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT | | Adapting the ALBERT model to biomedical and clinical corpora to obtain a new model named BioALBERT. | | ALBERT, BioALBERT | | BioALBERT has the same structure as ALBERT and addresses the shortcomings of BERT-based biomedical models. BioALBERT uses cross-layer parameter sharing and reduces 110 million parameters of the 12-layer BERTbase model to 31 million parameters while keeping the same number of layers and hidden units. BioALBERT uses sentence order prediction loss rather than BERT’s next sentence prediction. BioALBERT was initialized with ALBERT’s weights and then further trained on abstracts from PubMed, full-text articles of PMS and clinical notes. Sentence embeddings were used for tokenization by preprocessing each line as a sentence with a maximum length of 512 words. In total, 8 BioALBERT variants are created, these varying in the number of training sets and the amount of corpora used for training. The model was tested on a number of downstream BioNLP tasks, such as named entity recognition, relation extraction and document classification. In this literature review we focus on the document classification task as it is most topical. In this document classification task, multiple labels from texts are predicted. The hallmarks of Cancer dataset is used. On this task, BioALBERT is compared to the state-of-the-art at the time, which is ALBERT. | | BioALBERT was initialized using the ALBERT model, which is trained on 2.5 billion English Wikipedia words and 800 million BookCorpus words. BioALBERT is furtherly trained on 4.5 billion PubMed words, 13.5 billion PMC words and 500 million MIMIC-III words. For the document classification task the hallmark of cancer (HoC) dataset is used, which has a training size of 1.108, a development size of 157 and a test size of 315. | F1-scores on HoC document classification task: SOTA (ALBERT): 0.873, best performing BioALBERT model (trained on English Wikipedia, BooksCorpus and PubMed data with 200.000 training steps): **0.8792** | |
| Comments: In this paper, a helpful alteration to ALBERT is proposed for the biomedical domain. This is especially helpful for my research as it takes place in my domain. The model also improves upon ALBERT on tasks in the biomedical domain. I believe it would be beneficial to finetune this model on my lifestyle classification task in order to be able to compare performance to my BERT model. | | | | | | | | | | | | |
| 62 | | ConBERT: A Concatenation of Bidirectional Transformers for Standardization of Operative Reports from Electronic Medical Records | | Proposing a concatenation of BERT models (ConBERT) for predicting ICD-9 codes using operative reports and diagnoses recorded in free text. | | BERTbase, UmlsBERT, MedicalBERT, BioBERT, Medical Character BERT | | On the input data, four types of preprocessing were applied, these being data cleansing, removal of stop words, combining operative report and diagnosis and removal of duplicate data. For ConBERT, BERT-based embedding is applied. Instead of the word-based embedding of BERT a concatenation of both word-based and character-based embedding vectors is applied with the goal of boosting performance. The concatenated embedding vector was converted into a probability vector through a fully connected layer and an output sigmoid layer which outputs probabilities. Finetuning is done by adding a classification layer at the end. Training and validation was done using 10-fold cross validation on the input dataset. Ratio training test data is 80-20. | | The data stems from two Korean university hospitals and contains all surgical records of patient who underwent surgery between 2009 and 2021. The independent dataset sizes were 45.211 and 35.862 records respectively. The datasets are labelled with ICD-9 codes. | ICD-9 prediction macro F1-scores singular models: BERTbase: 0.321, UmlsBERT: **0.3383**, MedicalBERT (word-level embedding): 0.3333, BioBERT: 0.3369, MedicalBERT (character-level embedding): 0.2965.  Aggregated models: Concatenation of word and character-level embeddings on MedicalBERT: 0.3578, UmlsBERT + MedicalBERT (character-level embedding): **0.3643**, BioBERT + MedicalBERT (character-level embedding): 0.3637 | |
| Comments: This paper includes a model that was concatenated from two other BERT models and shows that such an approach could improve performance. I will take model concatenation into account, but will not consider this particular ConBERT model as it is specialized towards operative reports. | | | | | | | | | | | | |
| 65 | | Med-BERT: pretrained contextualized embeddings on largescale structured electronic health records for disease prediction | | Adapting the BERT framework to the structured EHR domain | | Med-BERT | | Med-BERT is trained on structured diagnosis data coded using ICD codes. It differs from BERT in that it uses code embeddings to represent each clinical code, visit embeddings to differentiate visits and a transformer structure to capture intercorrelations between codes. Another pretraining task was designed which concerned prediction of prolonged length of stay in hospital (Prolonged LOS), this pretraining task replaces NSP. Med-BERT was evaluated by finetuning on two disease prediction tasks: the prediction of heart failure among diabetic patients and the prediction of onset of pancreatic cancer. Med-BERT is compared to other models by comparing performance directly and recording performance of putting the other models on top of it. Logistic regression and random forest are used as baselines. | | The pretraining cohort for Med-BERT consists of 28 million patients extracted from the Cerner Health Facts database. | Regarding the AUC values on both the heart failure prediction and the prediction of onset of pancreatic cancer Med-BERT with the Bi-GRU model on top performed the best among 13 checked models. The AUC was 85.39 on the first task and 82.23 on the second task. | |
| Comments: I can use this paper as proof that the alteration of embeddings compared to the original BERT is a more widespread technique to boost performance. Regarding this model, I will not consider it further as it was built with structured health records and it was not compared to other BERT models, only recurrent neural networks, making it hard to discern whether this model performs to a standard which is acceptable for my research. | | | | | | | | | | | | |
| 67 | | DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter | | Pretraining a general-purpose language model that is smaller than BERT and performs similarly named DistilBERT. | | BERTbase, DistilBERT | | Knowledge distillation is a compression technique where a compact model, the student, is trained to reproduce the behaviour of a larger model, the teacher. During training, the distillation loss is minimized. The final training objective is a linear combination of the distillation loss with the supervised training loss, here the masked language modelling loss. Furthermore, the cosine embedding loss is added which will tend to align the direction of the student and teacher hidden states vectors. DistilBERT has the same general architecture as BERT. The token-type embeddings and the pooler are removed and the number of layers is reduced by a factor of 2. DistilBERT is initialized by taking one of the two layers of BERT. DistilBERT is trained on the same corpus as the original BERT model. DistilBERT was evaluated on the GLUE benchmark and several downstream tasks such as IMDb sentiment classification and SQuAD v1.1. | | DistilBERT was trained on the same corpus as the original BERT model, this being a concatenation of English Wikipedia and BookCorpus. | Results on development set of GLUE benchmark (median of 5 runs with different seeds): ELMo: 0.687, BERTbase: **0.795**, DistilBERT: 0.77  Accuracy on IMDb classification: BERTbase: **0.9346**, DistilBERT: 0.9282  F1-scores on SQuAD v1.1: BERTbase: **0.885**, DistilBERT: 0.858. DistilBERT with extra distillation during finetuning: 0.868 | |
| Comments: This model shows great promise as it is way smaller than BERT yet yields similar performance on classification tasks after finetuning. Whether I use the distillation proposed in this paper depends on the resources I have access to. If these resources severely restrict my ability to train a general BERT model it might be beneficial to utilize a distilled version like DistilBERT. | | | | | | | | | | | | |
| 77 | | MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records | | Pretraining a BERT model (MedRoBERTa.nl) on Dutch hospital notes and finetuning it on classifying sentences | | mBERT, BERTje, RobBERT, MedRoBERTa.nl | | In this study, it was decided to build models upon the RoBERTa architecture instead of the BERT architecture, in particular RobBERT as the source dataset is in Dutch, because of the lack of necessity for the NSP objective, the expectation that larger input sequences result in a better representation of long-term dependencies, a preference for a byte-level BPE tokenizer over a character-level BPE tokenizer and the expectation that a bigger batch size leads to better performance. In this study, both extending pretraining of the existing RobBERT model and pretraining a BERT model from scratch are considered and evaluated as both ways have advantages and disadvantages. Extending pretraining is done by initializing the RobBERT model and the domain-specific vocabulary and tokenizer, then freezing the model and training only the embedding lookup layer in order to adapt this layer to the domain-specific data. This ensures the model will link the embedding stored in the transformer layers of a specific token to the same token in the new vocabulary. Next, the entire model is trained for 2 epochs. For the pretraining from scratch they start with random weights for the model as weights from an already finished Dutch model might not be suitable to Dutch medical texts because of the different terminology. Evaluation is done by testing the models on an annotated sample from the dataset. Furthermore, an odd-one-out similarity test is created where the models are presented with three sentences and the task is to find the sentence that least fits the other two. Extrinsic evaluation is also done by finetuning the models on one of the ICF-classification datasets. | | MedRoBERTa.nl was trained on hospital notes from two Dutch hospitals, in total there were around 11 million notes, 13 GB. | Accuracy per model on test set of annotated data: mBERT: 0.57, BERTje: 0.58, RobBERT: 0.57, MedRoBERTa.nl pretrained from scratch: **0.65**, RobBERT extended: 0.58  For the ICF extrinsic evaluation the pretrained model from scratch outperformed the other models on two of the four types. | |
| Comments: This paper is extremely relevant to my research. The usage of Dutch clinical notes is exactly the same as my research. Furthermore, every decision that was made is elaborated extensively, helping me a lot to consider which choices need to be made. I also need to choose between RoBERTa and BERT architectures. I can use this model both as a source for ideas for my model and as a means of comparing in evaluating my model. I could also use this model as a base, just like RobBERT is used as a base here and pretraining is done on top of that. But then, pretraining from scratch outperformed extending RobBERT, so that is also a valuable insight due to the extreme similarities of the objective of their study and mine. I will definitely compare pretraining BERT models from scratch and extending another model’s pretraining and this model is the strongest candidate for the latter. | | | | | | | | | | | | |
| 80 | | Text-based classification of interviews for mental health - juxtaposing the state of the art | | Introducing belabBERT, a new Dutch language model extending the RoBERTa architecture. The goal is to extend the framework to hybrid text- and audio-based classification. | | RobBERT, belabBERT | | The aim is to create a model that is able to perform classification based on only text. After this, a hybrid model that also incorporates audio classification is created. For preprocessing, the data was transformed to flat text. belabBERT differs from RobBERT in that is not trained on a shuffled version of the OSCAR Web crawl corpus, rather an unshuffled version, as this was prophesized in a meta-analysis study to be performing better. By using this non-shuffled version the sentence order of the corpus is preserved, which could enable belabBERT to learn long range syntactic dependencies. Text was preprocessed by fuzzy deduplication (removing lines with over 90% overlap), removing non-textual data and excluding lines longer than 2.000 words. The same training parameters as RoBERTa were used for belabBERT. Finetuning was done for the classification of text input. The audio model was created using a simple architecture of three layers. The hybrid model combines the models. In this literature review, I focus on the text classification model as it is most topical. | | belabBERT was trained on 32 GB of preprocessed Dutch text. 10% of the text was held out as validation set. Finetuning was done on 141 transcribed interviews, which were split into training, development and test sets with ratio 80-10-10. | Accuracy on interview test set: belabBERT: **0.7568**, RobBERT: 0.6906 | |
| Comments: This paper is interesting as it concerns a Dutch model which improves upon RobBERT, which is considered to be the Dutch general text state-of-the-art. The model was trained on a relatively huge corpus of Dutch text, larger than BERTje and RobBERT. As the finetuning task is so small, it might be interesting to figure out how well the model is able to finetune on a Dutch clinical note lifestyle classification task, like the one in my research. This model is also a candidate to use as a basis for my model, as I could extend upon its pretraining by pretraining it furtherly on clinical nodes, as has happened with RobBERT and MedRoBERTa.nl. | | | | | | | | | | | | |
| 83 | | Traditional Chinese medicine clinical records classification with BERT and domain specific corpora | | Classifying Traditional Chinese Medicine (TCM) clinical records into 5 main disease categories, introducing new BERT variant TCM-BERT. | | TCM-BERT, BERT | | TCM clinical notes are regarded to be quite different from general domain texts, as they contain many TCM domain specific terms. For this reason, the pretrained BERT language model is finetuned on the TCM clinical corpus before finetuning it as a text classifier. Pretraining is largely the same as the BERT pretraining. For pretraining they use the BERTbaseChinese model. The NSP and MLM tasks are intact. TCM-BERT is compared to a number of representative deep learning models, such as PV-DBOW + SVM (a document embedding method with SVM classifier), ESA-PV-DBOW + SVM (a knowledge-based document embedding method with SVM classifier), Word Tf-idf + SVM, Char Tf-idf + SVM, CNN’s, Bi-LSTM’s and fastText models. Classifying is done based on grouping records into 5 disease categories. | | The benchmark TCM clinical records dataset was used, which contains 7.037 records. The training/test set ratio is 70-30. | Macro F1-score on test set: PV-DBOW + SVM: 0.7635, ESA-PV-DBOW + SVM: 0.812, Word TFIDF + SVM: 0.7564, Char TFIDF + SVM: 0.7386, Char CNN: 0.8411, Word CNNL 0.8027, Char Bi-LSTM: 0.5663, Char Bi-LSTM + Attention: 0.7986, Word Bi-LSTM: 0.5623, Char fastText: 0.7695, Char fastText (bigrams): 0.7955, Word fastText: 0.7342, BERT: 0.8708, TCM-BERT: **0.8867** | |
| Comments: This paper shows that finetuning a BERT model on a dataset and then finetuning it for a specific classification task could outperform just finetuning BERT on the classification task. This is something that also shows up in a lot of other studies covered in this literature review. It also shows BERT-models severely outperforming other NLP models which is also a recurring theme. I will not use this new BERT model furtherly, as it is finetuned on a task which is far removed from mine. | | | | | | | | | | | | |