tom recognition from colloquial narratives of patients. Informal, colloquial texts are a challenge to the state

and terminology than formal texts. Using text data from COVID-19-related online patient forums and medical papers, we conduct experiments comparing performance of BERT on normalized and de-normalized variants of the data. These experiments will shed light on the impact of various text characteristics on model performance, based on which design principles for domain-specific model training for NER tasks can be developed.

Keywords: Named Entity Recognition, BERT, patient narratives, symptoms

1 Introduction

Named entity recognition (NER) seeks to automatically extract terms representing entities (e.g., persons, organizations, locations, etc.) from documents. NER has long been an important task in natural language processing (NLP) applications. Recently, the introduction of the BERT (Bidirectional Encoder Representations from Transformers) model [2] has brought impressive progress in open-domain NLP tasks, including NER.

are

approach for this domain-specific NER task. From the design science perspective, we seek to develop a set of design artifacts, which will include not only effective, domain-specific NER methods for handling informal writings but also general design principles for such tasks. We start with identifying sources of impact on NER performance when dealing with colloquial writing, aiming to discover *what* affects performance and *how* we can improve it. Our research may help healthcare professionals, medical researchers, pharmaceutical manufacturers, and disease control organizations to mine valuable information and knowledge from text data for a variety of purposes, such as studying new epidemics or diseases, gathering patient reactions to treatments, documenting side effects of drugs, and monitoring spread of viruses (e.g., COVID-19).

In the following, we review the related work on open-domain BERT and NER in the healthcare and medical domain. We present our research design and preliminary results from symptom recognition in both patient narratives and medical literature. At the end21ci1 0 0 1 349.734220.000008

outperform the open-domain BERT. Alternatively, CT-BERT is a model trained on COVID-19 related Twitter messages and has achieved a 10-30% improvement over BERT on sentiment analysis of tweets [11].

Named Entity Recognition (NER) is an information extraction task for identifying and extracting entities of interest, such as persons (e.g., John Doe), organizations (e.g., the National Science Foundation), locations (e.g., Boston), date and time (e.g., July 4th, Saturday morning) from text documents. Traditional NER methods include lexicon-based string matching and pattern recognition, rule-based heuristics, statistical models, and classification in machine learning. The performance of word-context-based statistical and machine learning approaches, such as Conditional Random Fields (CRF) and LSTM (Long Short-Term Memory), depends heavily on feature engineering, in which part-of-speech tags, the position of the word in a sentence, and other word-based features are constructed and used as input to train and build the classifier. A large amount of prior NER studies focus on constructing features for classification [13].

The research on NER has benefited tremendously from the advent of BERT. Without having to depend on extensive feature engineering, the open-domain BERT has outperformed traditional NER approaches ([2, 6]). BERT-based models have been applied to extracting *domain-specific* entities such as medicine and treatment names in medical documents. For instance, BioBERT [4] improves the F1 score of NER in biomedical texts. Incorporating knowledge about biomedical entities (e.g., chemicals and proteins)

of early and open communication during the beginning of the pandemic, in which pa-