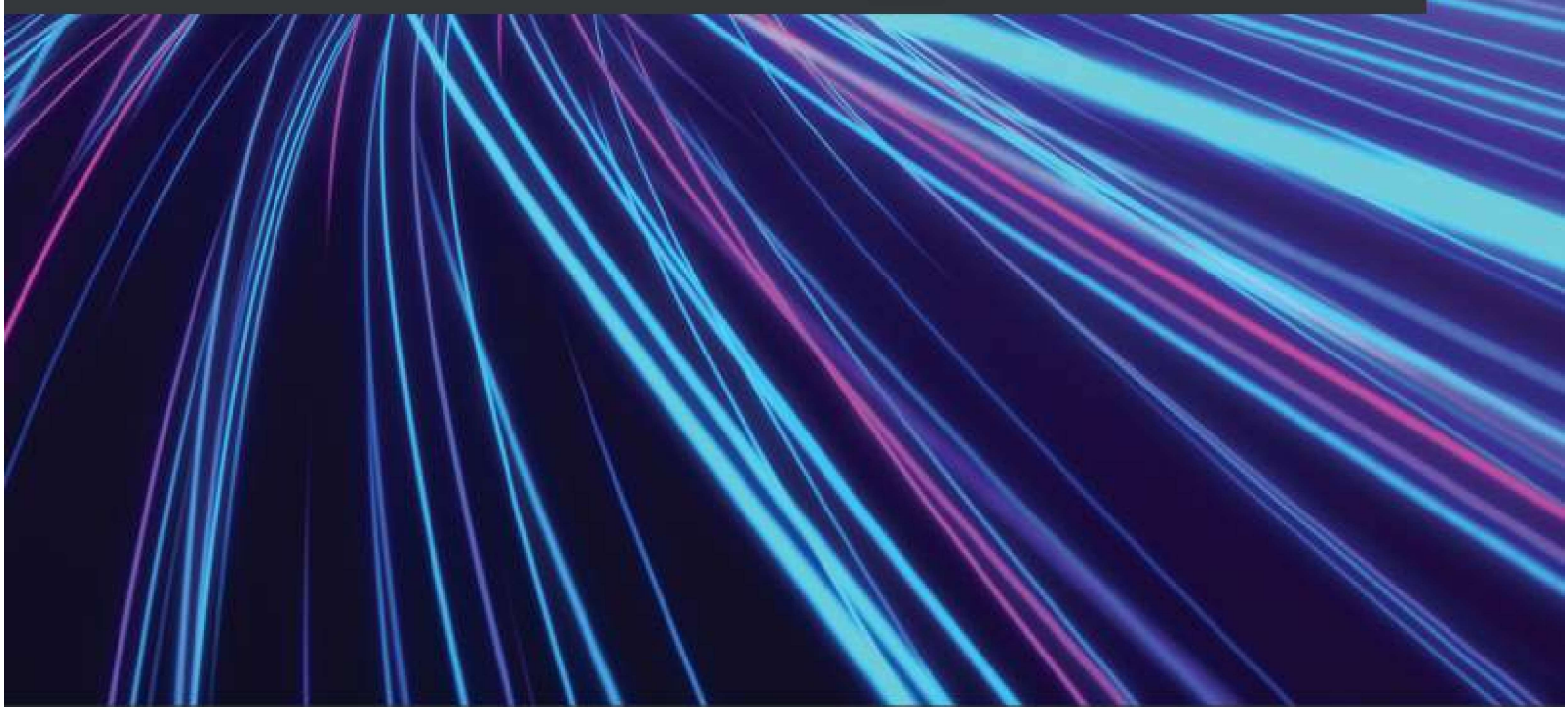
Getting Started with Google BERT

Build and train state-of-the-art natural language processing models using BERT



Sudharsan Ravichandiran

Getting Started with Google BERT

Build and train state-of-the-art natural language processing models using BERT

Sudharsan Ravichandiran

Packt

BIRMINGHAM - MUMBAI

Getting Started with Google BERT

Group Product Manager: Kunal Parikh
Publishing Product Manager: Devika Battike
Content Development Editor: Sean Lobo
Senior Editor: Roshan Kumar
Technical Editor: Manikandan Kurup
Copy Editor: Safis Editing
Project Coordinator: Aishwarya Mohan
Proofreader: Safis Editing
Indexer: Priyanka Dhadke
Production Designer: Prashant Ghare

First published: January 2021

Production reference: 1210121

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-83882-159-3

www.packt.com

To my adorable mom, Kasthuri, and to my beloved dad, Ravichandiran.

- Sudharsan Ravichandiran

Packt.com

Subscribe to our online digital library for full access to over 7,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Improve your learning with Skill Plans built especially for you
- Get a free eBook or video every month
- Fully searchable for easy access to vital information
- Copy and paste, print, and bookmark content

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.packt.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at customercare@packtpub.com for more details.

At www.packt.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on Packt books and eBooks.

About the author

Sudharsan Ravichandiran is a data scientist, researcher, and bestselling author. He completed his bachelor's in information technology at Anna University. His area of research focuses on practical implementations of deep learning and reinforcement learning, including natural language processing and computer vision. He is an open source contributor and loves answering questions on Stack Overflow. He also authored a best seller, *Hands-On Reinforcement Learning with Python*, published by Packt Publishing.

I would like to thank my most amazing parents and my brother, Karthikeyan, for inspiring and motivating me. I would like to thank the Packt team, Devika, Sean, and Kirti, for their great help. Without all of their support, it would have been impossible to complete this book.

About the reviewers

Dr. Armando Fandango creates AI-empowered products by leveraging reinforcement learning, deep learning, and distributed computing. Armando has provided thought leadership in diverse roles at small and large enterprises, including Accenture, Nike, Sonobi, and IBM, along with advising high-tech AI-based start-ups. Armando has authored several books, including *Mastering TensorFlow*, *TensorFlow Machine Learning Projects*, and *Python Data Analysis*, and has published research in international journals and presented his research at conferences. Dr. Armando's current research and product development interests lie in the areas of reinforcement learning, deep learning, edge AI, and AI in simulated and real environments (VR/XR/AR).

Ashwin Sreenivas is the cofounder and chief technology officer of Helia AI, a computer vision company that structures and understands the world's video. Prior to this, he was a deployment strategist at Palantir Technologies. Ashwin graduated in Phi Beta Kappa from Stanford University with a master's degree in artificial intelligence and a bachelor's degree in computer science.

Gabriel Bianconi is the founder of Scalar Research, an artificial intelligence and data science consulting firm. Past clients include start-ups backed by YCombinator and leading venture capital firms (for example, Scale AI, and Fandom), investment firms, and their portfolio companies (for example, the Two Sigma-backed insurance firm MGA), and large enterprises (for example, an industrial conglomerate in Asia, and a leading strategy consulting firm). Beyond consulting, Gabriel is a frequent speaker at major technology conferences and a reviewer on top academic conferences (for example, ICML) and AI textbooks. Previously, he received B.S. and M.S. degrees in computer science from Stanford University, where he conducted award-winning research in computer vision and deep learning.

Mani Kanteswara has a bachelor's and a master's in finance (tech) from BITS Pilani with over 10 years of strong technical expertise and statistical knowledge of analytics. He is currently working as a lead strategist with Google and has previously worked as a senior data scientist at WalmartLabs. He has worked in deep learning, computer vision, machine learning, and the natural language processing space building solutions/frameworks capable of solving different business problems and building algorithmic products. He has extensive expertise in solving problems in IoT, telematics, social media, the web, and the e-commerce space. He strongly believes that learning concepts with a practical implementation of the subject and exploring its application areas leads to a great foundation.

Packt is searching for authors like you

If you're interested in becoming an author for Packt, please visit `authors.packtpub.com` and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Table of Contents

Section 2 - Exploring BERT Variants

Preface

Bidirectional Encoder Representations from Transformers (**BERT**) has revolutionized the world of **natural language processing** (**NLP**) with promising results. This book is an introductory guide that will help you get to grips with Google's BERT architecture.

The book begins by giving you a detailed explanation of the transformer architecture and helps you understand how the encoder and decoder of the transformer work.

You'll get to grips with BERT and explore its architecture, along with discovering how the BERT model is pre-trained and how to use pre-trained BERT for downstream tasks by fine-tuning it. As you advance, you'll find out about different variants of BERT such as ALBERT, RoBERTa, ELECTRA, and SpanBERT, as well as look into BERT variants based on knowledge distillation, such as DistilBERT and TinyBERT. The book also teaches you about M-BERT, XLM, and XLM-R in detail. You'll then learn about Sentence-BERT, which is used for obtaining sentence representation. You will also see some domain-specific BERT models such as BioBERT and ClinicalBERT. At the end of the book, you will learn about an interesting variant of BERT called VideoBERT.

By the end of this BERT book, you'll be well versed in using BERT and its variants for performing practical NLP tasks.

Who this book is for

This book is for NLP professionals and data scientists looking to simplify NLP tasks to enable efficient language understanding using BERT. A basic understanding of NLP concepts and deep learning is required to get the most out of this book.

What this book covers

`Chapter 1`, *A Primer on Transformers*, explains the transformer model in detail. We will understand how the encoder and decoder of transformer work by looking at their components in detail.

`Chapter 2`, *Understanding the BERT model*, helps us to understand the BERT model. We will learn how the BERT model is pre-trained using **Masked Language Model** (**MLM**) and **Next Sentence Prediction** (**NSP**) tasks. We will also learn several interesting subword tokenization algorithms.

`Chapter 3`, *Getting Hands-On with BERT*, explains how to use the pre-trained BERT model. We will learn how to extract contextual sentences and word embeddings using the pre-trained BERT model. We will also learn how to fine-tune the pre-trained BERT for downstream tasks such as question-answering, text classification, and more.

`Chapter 4`, *BERT Variants I – ALBERT, RoBERTa, ELECTRA, and SpanBERT*, explains several variants of BERT. We will learn how BERT variants differ from BERT and how they are useful in detail.

`Chapter 5`, *BERT Variants II – Based on Knowledge Distillation*, deals with BERT models based on distillation, such as DistilBERT and TinyBERT. We will also learn how to transfer knowledge from a pre-trained BERT model to a simple neural network.

`Chapter 6`, *Exploring BERTSUM for Text Summarization*, explains how to fine-tune the pre-trained BERT model for a text summarization task. We will understand how to fine-tune BERT for extractive summarization and abstractive summarization in detail.

`Chapter 7`, *Applying BERT to Other Languages*, deals with applying BERT to languages other than English. We will learn about the effectiveness of multilingual BERT in detail. We will also explore several cross-lingual models such as XLM and XLM-R.

`Chapter 8`, *Exploring Sentence and Domain-Specific BERT*, explains Sentence-BERT, which is used to obtain the sentence representation. We will also learn how to use the pre-trained Sentence-BERT model. Along with this, we will also explore domain-specific BERT models such as ClinicalBERT and BioBERT.

`Chapter 9`, *Working with VideoBERT, BART, and More*, deals with an interesting type of BERT called VideoBERT. We will also learn about a model called BART in detail. We will also explore two popular libraries known as ktrain and bert-as-service.

To get the most out of this book

To get the most out of the book, run all the code provided in the book using Google Colab.

Software/Hardware requirements	**Operating System**
Google Colab / Python 3.x	Windows/macOS/Linux

Download the example code files

You can download the example code files for this book from GitHub at `https://github.com/PacktPublishing/Getting-Started-with-Google-BERT`. In case there's an update to the code, it will be updated on the existing GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: `https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf`.

Conventions used

There are a number of text conventions used throughout this book.

`CodeInText`: Indicates code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles. Here is an example: "We will set `maxlen` to `100` and `max_features` to `100000`."

A block of code is set as follows:

```
(x_train, y_train), (x_test, y_test), preproc = \
text.texts_from_df(train_df = df,
                   text_column = 'reviewText',
                   label_columns=['sentiment'],
                   maxlen=100,
                   max_features=100000,
                   preprocess_mode='bert',
                   val_pct=0.1)
```

Bold: Indicates a new term, an important word, or words that you see onscreen. For example, words in menus or dialog boxes appear in the text like this. Here is an example: "Select **System info** from the **Administration** panel."

Warnings or important notes appear like this.

Tips and tricks appear like this.

Get in touch

Feedback from our readers is always welcome.

General feedback: If you have questions about any aspect of this book, mention the book title in the subject of your message and email us at `customercare@packtpub.com`.

Errata: Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you have found a mistake in this book, we would be grateful if you would report this to us. Please visit `www.packtpub.com/support/errata`, selecting your book, clicking on the Errata Submission Form link, and entering the details.

Piracy: If you come across any illegal copies of our works in any form on the Internet, we would be grateful if you would provide us with the location address or website name. Please contact us at `copyright@packt.com` with a link to the material.

If you are interested in becoming an author: If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, please visit `authors.packtpub.com`.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions, we at Packt can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about Packt, please visit `packt.com`.

1
Section 1 - Starting Off with BERT

In this section, we will familiarize ourselves with BERT. First, we will understand how the transformer works, and then we will explore BERT in detail. We will also get hands-on with BERT and learn how to use the pre-trained BERT model.

The following chapters are included in this section:

- `Chapter 1`, *A Primer on Transformers*
- `Chapter 2`, *Understanding the BERT Model*
- `Chapter 3`, *Getting Hands–On with BERT*