

# Cambridge Elements

Elements in Quantitative and Computational Methods  
for the Social Sciences

edited by  
R. Michael Alvarez  
*California Institute of Technology*  
Nathaniel Beck  
*New York University*

## TEXT ANALYSIS IN PYTHON FOR SOCIAL SCIENTISTS

***Prediction and Classification***

Dirk Hovy  
*Bocconi University*



CAMBRIDGE  
UNIVERSITY PRESS

Cambridge University Press  
978-1-108-95850-9 — Text Analysis in Python for Social Scientists  
Dirk Hovy  
Frontmatter  
[More Information](#)

---

**CAMBRIDGE**  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom  
One Liberty Plaza, 20th Floor, New York, NY 10006, USA  
477 Williamstown Road, Port Melbourne, VIC 3207, Australia  
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,  
New Delhi – 110025, India  
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.  
It furthers the University's mission by disseminating knowledge in the pursuit of  
education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781108958509](http://www.cambridge.org/9781108958509)  
DOI: 10.1017/9781108960885

© Dirk Hovy 2022

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

First published 2022

*A catalogue record for this publication is available from the British Library.*

ISBN 978-1-108-95850-9 Paperback  
ISSN 2398-4023 (online)  
ISSN 2514-3794 (print)

Cambridge University Press has no responsibility for the persistence or accuracy of  
URLs for external or third-party internet websites referred to in this publication  
and does not guarantee that any content on such websites is, or will remain,  
accurate or appropriate.

## Text Analysis in Python for Social Scientists

Prediction and Classification

Elements in Quantitative and Computational Methods for the Social Sciences

DOI: 10.1017/9781108960885  
First published online: February 2022

Dirk Hovy  
*Bocconi University*

**Author for correspondence:** Dirk Hovy, [dirk.hovy@unibocconi.it](mailto:dirk.hovy@unibocconi.it)

**Abstract:** Text contains a wealth of information about a wide variety of sociocultural constructs. Automated prediction methods can infer these quantities (sentiment analysis is probably the most well-known application). However, there is virtually no limit to the kinds of things we can predict from the text: power, trust, and misogyny are all signaled in language. These algorithms easily scale to corpus sizes infeasible for manual analysis. Prediction algorithms have become steadily more powerful, especially with the advent of neural network methods. However, applying these techniques usually requires profound programming knowledge and machine learning expertise. As a result, many social scientists do not apply them. This Element provides the working social scientist with an overview of the most common methods for text classification, an intuition of their applicability, and Python code to execute them. It covers both the ethical foundations of such work as well as the emerging potential of neural network methods.

**Keywords:** text analysis, natural language processing, computational linguistics, classification, prediction

**JEL classifications:** A12, B34, C56, D78, E90

© Dirk Hovy 2022

ISBNs: 9781108958509 (PB), 9781108960885 (OC)  
ISSNs: 2398-4023 (online), 2514-3794 (print)

## Contents

Introduction	1
Background: Classification and Prediction	2
1 Ethics, Fairness, and Bias	3
Prediction: Using Patterns in the Data	11
2 Classification	11
3 Text as Input	17
4 Labels	20
5 Train-Dev-Test	22
6 Performance Metrics	25
7 Comparison and Significance Testing	29
8 Overfitting and Regularization	33
9 Model Selection and Other Classifiers	36
10 Model Bias	40
11 Feature Selection	41
12 Structured Prediction	45
Neural Networks	54
13 Background of Neural Networks	54
14 Neural Architectures and Models	70
References	83