



Bachelor- oder Masterarbeit zum Thema

**„Freitextklassifikation von offenen Antworten zur Verbesserung des
Mikrozensus“**

im Landesbetrieb

Information und Technik Nordrhein-Westfalen (IT.NRW) – Statistisches Landesamt

Über uns: Die amtliche Statistik hat den gesetzlichen Auftrag zur Erhebung, Aufbereitung und Veröffentlichung von Daten. Als Statistisches Landesamt für Nordrhein-Westfalen nimmt der Geschäftsbereich Statistik des [Landesbetriebes Information und Technik Nordrhein-Westfalen \(IT.NRW\)](#) diese Aufgabe wahr.

Über das Anwendungsgebiet: Der Mikrozensus ist eine jährliche Befragung von privaten Haushalten, die seit 1957 durchgeführt wird. Die erhobenen Daten bieten vielfältige Erkenntnisse zu den Lebens- und Arbeitsbedingungen und der Struktur der Bevölkerung. In Nordrhein-Westfalen werden jährlich rund 14.000 Auswahlbezirke nach einem mathematischen Zufallsverfahren ausgewählt und befragt.

Thema der Arbeit: Im Rahmen des Mikrozensus werden mittels Online-Formularen Daten erhoben. Am Ende des Online-Fragebogens gibt es ein Freitextfeld für zusätzliche Anmerkungen und Hinweise. Das Freitextfeld wird von den Auskunftgebenden für sachdienliche Hinweise, die Verbesserungen im Erhebungsprozess oder inhaltliche Klärungen ermöglichen, genutzt. Allerdings erfolgen über das Freitextfeld auch unsachliche Anmerkungen und allgemeine Beschwerden. Dies erschwert die Auswertung, da auf sachdienliche Anmerkungen schwerer reagiert werden kann.

Ziel der Arbeit: Ziel des Projekts ist die Klassifikation der Freitextantworten in Angaben, auf die reagiert werden muss, und Angaben, auf die *nicht* reagiert werden muss. Damit ließen sich perspektivisch Effizienz und Qualität in der Erhebung des Mikrozensus verbessern. Für das Erhebungsjahr 2025 gingen etwa 10.000 Freitextantworten ein. Eine rein manuelle Bearbeitung kann bei dieser Menge nicht gewährleistet werden.

Die Klassifikation von Texten eignet sich hervorragend für diverse Verfahren des maschinellen Lernens oder den Einsatz von LLMs.

Ihre Voraussetzungen:

- Sie befinden sich im Bachelor- oder Masterstudium im Bereich Statistik, Data Science oder Econometrics.
- Sie haben Interesse an der amtlichen Statistik und deren Fragestellungen.
- Sie verfügen über ausgeprägte Kenntnisse in gängigen Programmiersprachen wie R oder Python.



- Sie sind bereit, ggf. Ergebnisse bei Forschungstreffen und bei IT.NRW (online) zu präsentieren und ggf. zu wissenschaftlichen Publikationen beizutragen

Was wir bieten:

- Das Projekt kann wahlweise als Bachelor- oder Masterarbeit durchgeführt werden.
- Die Arbeit kann teilweise im Rahmen einer vergüteten Werksstudierendentätigkeit bei IT.NRW geleistet werden. Während dieser Zeit erhalten Sie Zugang zu den für die Arbeit relevanten Daten.
- Es erfolgt ein regelmäßiger Austausch mit den Fachexpertinnen und -experten bei IT.NRW.
- Möglichkeit zur Arbeit von zu Hause.

Haben Sie Interesse? Wir freuen uns über Ihre Bewerbung!

Inhaltliche und organisatorische Rückfragen beantworten Ihnen gern:

Prof. Dr. Katja Ickstadt
0231 755-3111
ickstadt@statistik.tu-dortmund.de

Dr. Christian Borgs
Referent Mathematisch-statistische Methoden und Experimentelle Statistik
0211 9449-2514
Christian.Borgs@it.nrw.de

Referenzen:

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8), 966-974.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3), 1-40.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial intelligence review*, 52(1), 273-292.