„Wir sind keine Zweck-WG!“
Analyse von WG-Anzeigen mit Topic Modeling

Elisa Gilbert
3780947

13. März 2025

Introduction to Digital Humanities
Universität Leipzig

Manuel Burghardt, Thomas Efer & Andreas Niekler

Wintersemester 2024/2025

Repository: https://github.com/chill-bird/topic_modeling_wggesucht

1 Einleitung

Das Fach der Digital Humanities (DH) schlägt die Brücke zwischen den klassischen Geisteswissenschaften und der Informatik. Sie verbinden Inhalte und Methoden beider Domänen und haben hohen interdisziplinären Charakter.

Der Forschungsgegenstand der DH umfasst geisteswissenschaftliche Themenbereiche, wie etwa den Literatur- und Kulturwissenschaften, der Linguistik oder Musikwissenschaft. Die Methodik der DH bedient sich hingegen informatischer Werkzeuge, insbesondere zur computergestützten Analyse und Visualisierung von Daten.

Die Ursprünge der Digital Humanities werden häufig mit dem Projekt *Index Thomisticus* in den 1940er Jahren verortet. Roberto Busa erfasste dabei die 118 Werke von Thomas von Aquin durch statistische Listen und Konkordanzen (Zemanek, 1977). Seitdem kann die zunehmende Digitalisierung in unserem Informationszeitalter als Wegbereiter und Katalysator der DH betrachtet werden. Die schnelle Verarbeitung großer Datenmengen durch moderne Computer sowie der Zugang zu offenen Daten, spezialisierten Werkzeugen und nativ digitalen Quellen bereichern die Disziplin kontinuierlich. Die DH eröffnen somit eine zeitgemäße, erweiterte Perspektive auf geisteswissenschaftliche Inhalte.

Die genaue Definition und Abgrenzung der Digital Humanities ist Gegenstand einer langjährigen wissenschaftlichen Debatte. Roth (2019) identifiziert in ihrer Übersichtsarbeit drei zentrale Aspekte: *Numerical Humanities* (bzw. *Computational Humanities*), *Digitalized Humanities* und *Humanities of the Digital*. Hiernach liegen die Themen- und Verfahrensschwerpunkte der DH in der mathematischen Abstraktion und formalen Modellierung, digitalisierten geisteswissenschaftlicher Ressourcen und computervermittelten Interaktionen und Online-Communities.

Die Digital Humanities bieten eine Vielzahl von Methoden zur computergestützten Analyse geisteswissenschaftlicher Fragestellungen. Zur Untersuchung sozialer und linguistischer Muster stellt die Methode des Topic Modelings eine beliebte Technik innerhalb der DH dar (Chen et al., 2023). Hierbei werden latente Themen in Textkorpora automatisiert identifiziert. Diese Methode wird im vorliegenden Projekt zur Analyse von Online-Anzeigen für Wohnungsgemeinschaften (WGs) genutzt. Dabei stehen insbesondere die Aspekte der „Numerical Humanities“ und der „Digitalized Humanities“ im Fokus: Die latenten Themen einer textuellen Online-Ressource werden mithilfe eines stochastischen Modells (Latent Dirichlet Allocation, LDA) abgebildet und qualitativ untersucht. Ein wesentlicher geisteswissenschaftlicher Aspekt dieses Projekts liegt in der sozialwissenschaftlichen und sprachwissenschaftlichen Analyse der WG-Anzeigen. Die Untersuchung der latenten Themen in den Beschreibungstexten ermöglicht Rückschlüsse auf Aspekte des gemeinschaftlichen Wohnens und kommunikative Konventionen in digitalen Wohnungsanzeigen. Noch dazu ermöglicht das vorliegende Projekt die Untersuchung möglicher thematischer Themenverlagerungen zwischen dem Zusammenleben in FLINTA*-WGs und Non-FLINTA*-WGs. Somit bewegt sich das Projekt interdisziplinär an der Schnittstelle zwischen Linguistik, Soziologie und Kulturwissenschaften.

2 Forschungsagenda

In diesem Projekt werden WG-Anzeigen der Online-Plattform *WG-gesucht* untersucht. Hierbei sollen zwei Forschungsfragen adressiert werden:

1. Welche Themen sind in den WG-Beschreibungstexten am stärksten vertreten?
2. Gibt es Unterschiede zwischen in der Themenverteilung von FLINTA*- und Non-FLINTA*-WGs?

Als Corpus des Projekts dienen die Beschreibungstexte der WG-Anzeigen. Üblicherweise werden diese Texte von den WGs selbst verfasst, um neue Mitbewohnende zu suchen. Diese Praxis erlaubt es daher, die Beschreibungen heranzuziehen, um Rückschlüsse auf die jeweiligen WGs, aber auch auf kommunikative Konventionen auf der Plattform ziehen. Die Texte enthalten Beschreibungen der WG und ihrer Mitglieder, Erwartungen an Bewerbende sowie Details zur Wohnung und dem angebotenen Zimmer. Bei der Analyse der Texte wird erwartet, dass Themenschwerpunkte identifiziert werden können, die digitale WG-Anzeigen charakterisieren und die möglicherweise Rückschlüsse auf das Zusammenleben in den Wohnungsgemeinschaften zulassen.

FLINTA*-WG-Anzeigen sind ein etablierter Bestandteil auf der Online-Plattform *WG-gesucht*. Das Akronym FLINTA* umfasst Frauen, Lesben, intergeschlechtliche, nichtbinäre, transgeschlechtliche und agender Personen. In dieser Arbeit ist der FLINTA*-Begriff gewählt worden, um eine inklusivere Gender-Perspektive mit Fokus auf der Selbstwahrnehmung zu wählen, die außerhalb eines binären biologistischen Geschlechterverständnisses verstanden werden soll. Die Rolle von Gender in Wohnungsgemeinschaften ist in der Forschung bislang unterrepräsentiert: Diehl et al. (2013) und Woo et al. (2019) finden zwar Belege dafür, dass all-female Wohnungsgemeinschaften tendenziell von Bewerbenden bevorzugt werden, jedoch merken beide Studien an, dass qualitative Untersuchungen zu dieser Thematik fehlen. Um sich diesem qualitativen Aspekt der Fragestellung anzunähern, werden die Themenverteilungen zwischen den FLINTA*-WGs und Non-FLINTA*-WGs untersucht. Hierbei wird überprüft, ob eine möglicherweise unterschiedliche Themenverteilung zwischen den beiden Gruppen Aufschluss darüber geben könnte, warum all-female WGs von Bewerbenden als attraktiver wahrgenommen werden.

3 Daten-Überblick

Der Datensatz umfasst WG-Anzeigen der Plattform *WG-gesucht*, die über einen Zeitraum von 60 Tagen (23. November 2023 – 21. Januar 2024) erhoben wurden. Er enthält er $n = 620$ Anzeigen für WG-Inserate in der Stadt Tübingen von der URL www.wg-gesucht.de/wg-zimmer-in-Tuebingen.127.0.1.0.html. Nach der Entfernung der Wörtern, die in weniger als 1% der Dokumente auftauchen, verbleiben noch $n = 527$ Dokumente im finalen Datensatz. Auf der Plattform können WGs Anzeigen schalten, um neue Mitbewohnende zu suchen. In der Regel werden diese von einem oder mehreren Bewohnenden der aktuellen

Konstellation einer Wohnungsgemeinschaft erstellt. Insbesondere für Studierende ist *WG-gesucht* eine verbreitete Möglichkeit, um eine Wohnungsgemeinschaft zu finden. Die Daten wurden im Rahmen eines vorherigen Universitätsprojekts erhoben.

Jeder Eintrag entspricht hierbei einer Anzeige auf der Plattform. Jede Anzeige enthält vier Beschreibungstexte der Kategorien „Zimmer“, „Lage“, „WG-Leben“ und „Sonstiges“. Die Beschreibungstexte der Kategorien „Lage“ und „Zimmer“ wurden in der Analyse nicht berücksichtigt, da hier vor allem die Entfernung zu den nächstgelegenen Einkaufsmöglichkeiten oder die Anbindung an den öffentlichen Nahverkehr sowie das Zimmer selbst beschrieben wird und aus diesen Gründen als inhaltlich wenig bedeutsam eingeschätzt wurden. Hingegen soll in den Beschreibungstexten der Kategorie „WG-Leben“ das gemeinsame Leben in der WG beschrieben sowie die Mitbewohnenden vorgestellt werden und in der Kategorie „Sonstiges“ weitere Angaben zur gesuchten Person gegeben werden. Die Beschreibungstexte sind Freitexte mit variabler Länge. Im Durchschnitt enthalten die konkatenierten Beschreibungstexte $\mu = 174.9$ Wörter pro Anzeige und reichen von der kürzesten Anzeige mit 3 Wörtern bis zur längsten mit 1165 Wörtern.

Neben den Beschreibungstexten sind weitere Metadaten für jede Anzeige verfügbar, wie etwa der Zimmerpreis, die Zimmergröße oder auch die eigens vergebenen Merkmale der WG (zum Beispiel „Berufstätigen-WG“ oder „vegetarisch/vegan“). Für den Schwerpunkt dieser Arbeit wurde die Information, ob es sich bei einer WG um eine FLINTA*-WG handelt, aus den Metadaten abgeleitet: Hierbei wurden WGs als FLINTA*-WGs kategorisiert, die in der aktuellen WG-Konstellation ausschließlich die Genderangaben *weiblich* oder *divers* vergeben haben und zusätzlich ein neues WG-Mitglied suchen, das entweder *weiblich* oder *divers* angibt. Aus diesen Informationen ergibt sich ein Verhältnis von 11.57% FLINTA*-WGs ($n = 61$ Anzeigen) und 88.43% Non-FLINTA*-WGs ($n = 466$ Anzeigen).

Bei den Ergebnissen der vorliegenden Projektarbeit sollten einige Limitationen des Datensatzes berücksichtigt werden. Mit insgesamt $n = 527$ Dokumenten mit pro Dokument durchschnittlich $\mu = 174.9$ Wörtern ist die Datenbasis relativ klein im Vergleich zu anderen Studien. Dies könnte die Generalisierbarkeit der Ergebnisse auf größere, heterogene Datensätze einschränken. Darüber hinaus ist das Datenset geografisch auf die Stadt Tübingen beschränkt und deckt mit einem Erhebungszeitraum von 60 Tagen nur einen kurzen Zeitraum ab. Diese Einschränkungen bedeuten, dass die Ergebnisse möglicherweise nicht auf andere Städte oder zeitlich größere oder abweichende Intervalle (wie etwa der Semesterbeginn der lokalen Hochschulen) übertragen werden können. Spezifisch für die Stadt Tübingen ergibt sich noch dazu eine mögliche Verzerrung der vertretenen WG-Mitglieder: Durch einen überdurchschnittlich hohen Bevölkerungsanteil von Studierenden könnten Berufstätige und Auszubildende innerhalb der WGs unterrepräsentiert sein („Tübingen“, 2025). Ein weiteres Problem ergibt sich aus der geringen Anzahl an verfügbaren Daten für die Gruppe der FLINTA*-WGs. Hinzu kommt, dass die Kategorisierung der WGs in FLINTA* und Non-FLINTA* manuell erfolgte und nicht auf einer Selbstklassifikation der WGs basiert.

4 Methode: Latent Dirichlet Allocation

Die zentrale Methode der Projektarbeit ist die Latent Dirichlet Allocation (LDA) von Blei et al. (2003), einer Methode des Topic Modelings.

Das Ziel des Topic Modelings ist es, auf automatisierte Art latente Themen aus einem Corpus zu extrahieren. Die Methode unterliegt der Annahme, dass sich die Bedeutung eines Wortes aus dem Kontext seiner Nachbarwörter ableiten lässt. Themen werden dabei als Muster gemeinsam auftretender Wörter innerhalb eines Textes erfasst (Jockers, 2013).

In der LDA besteht ein Corpus aus einer Menge von M Dokumenten (zum Beispiel Textabschnitten). Das Vokabular V ist die Menge aller Wörter des gesamten Corpus. Wörter werden so unter Berücksichtigung ihres Indexes innerhalb des Vokabulars (durch One-Hot-Encodings, beziehungsweise -Vektoren) modelliert. Jedes Dokument entspricht einer Sequenz von Wörtern $w = (w_1, w_2, ..., w_N)$, wobei w_N das N-te Wort innerhalb der Sequenz ist. (Blei et al., 2003).

Die LDA geht davon aus, dass jedes Dokument eine Mischung aus mehreren latenten Themen ist. Jedes Thema wird durch eine statistisch bedeutsame Verteilung über Wörter charakterisiert. Durch die Analyse der Häufigkeit und Konkurrenzen von Wörtern in den Texten wird versucht, die unterliegenden Themen der Dokumentsammlung zu identifizieren.

Die Methode implementiert ein Bayessches Modell, das zwei Zufallsvariablen θ und ϕ mithilfe von Statistischem Lernen schätzt beziehungsweise lernt. θ ist die Themenverteilung der Dokumente und ϕ die Wortverteilung der Themen. Dazu werden die Parameter K für die Anzahl der Themen sowie α und β für die globalen A-Priori-Verteilung festgelegt. Der Parameter α legt dabei die Verteilung der Themen pro Dokument fest, β die Verteilung der Wörter pro Thema (Blei et al., 2003).

Im Rahmen dieses Projekts sollten methodische Einschränkungen beachtet werden. Zuerst ist dazu die Größe des Datensatzes anzuführen. Der Datensatz umfasst lediglich $n = 527$ Dokumente mit durchschnittlich geringer Wortanzahl. Verglichen mit der Corpusgröße in anderen Forschungsarbeiten ist dies eher gering einzuschätzen (Blei et al. (2003) beispielsweise wendet in seiner Arbeit die Methode auf Datensätze mit 5000 bis 6100 Dokumenten an). Zusätzlich sind nur 61 der 527 Dokumente der Gruppe der FLINTA*-WGs zuzuordnen, was eine Berechnung zweier separater Modelle nicht umsetzbar macht. Ein weiterer Nachteil der Methode sind bilinguale Texte. Worte derselben Bedeutung in zwei Sprachen können mit dem LDA-Verfahren nicht ohne Weiteres aufeinander abgebildet werden und führen noch dazu bereits bei der Lemmatisierung zu Fehlern. Trotzdem wurden englische Texte für diese Arbeit nicht entfernt, da einige WGs auch Beschreibungen in beiden Sprachen (Deutsch und Englisch) gewählt haben und diese Daten dadurch fälschlicherweise entfernt würden. Weiterhin ist die Festlegung des Parameters K (Anzahl der Themen) als methodische Limitation zu benennen. Die Festlegung kann stark beeinflussen, welche Muster entdeckt werden: Eine zu hohe oder zu niedrige Themenanzahl kann relevante Inhalte überlagern oder wichtige Unterschiede unterschlagen.

5 Verwandte Arbeiten

Topic Modeling sowie im Speziellen LDA ist eine verbreitete Methode in den DH, bei denen Forschungsgegenstand häufig große textuelle Corpora sind (Blei, 2012). Im Folgenden sollen drei Studien vorgestellt werden, die diese Methode in unterschiedlichen Kontexten angewendet haben, um die Anwendungsbreite der Methode zu illustrieren.

5.1 Combining CDA and Topic Modeling: Analyzing Discursive Connections between Islamophobia and Anti-Feminism on an Online Forum (Törnberg & Törnberg, 2016)

Die Forschenden untersuchen die diskursiven Verbindungen zwischen Islamfeindlichkeit und Antifeminismus in Online-Foren. Ein Korpus von 12,796 Beiträgen aus dem schwedischen Forum *Flashback* wird mit einer Mixed-Methods-Strategie analysiert. Im ersten Schritt werden mithilfe von LDA Diskursive Felder herausgearbeitet, mit denen die Textpassagen klassifiziert werden. Nachfolgend werden die Dokumente betrachtet, die beide Aspekte (Antifeminismus und Islamfeindlichkeit) gleichzeitig betreffen. Diese Textpassagen werden anschließend mithilfe von Critical Discourse Analysis (CDA) analysiert. Hauptergebnisse der Studie sind, dass das Thema Feminismus im Zusammenhang mit Islamfeindlichkeit dazu genutzt wird, die muslimische Glaubensgemeinschaft abzuwerten, da sie Frauen unterdrücke und daher der schwedischen Kultur unterlegen sei. Trotzdem wird von denselben Usern der Feminismus als Gefährdung der schwedischen Kultur und ihrer Normen dargestellt, wodurch sich beide Argumentationen gegenseitig widersprechen.

5.2 The evolution of Airbnb research: A systematic literature review using structural topic modeling (Ding et al., 2023)

Die Studie untersucht die Entwicklung der wissenschaftlichen Forschung zu Airbnb, einem Online-Portal zur Buchung und Vermietung von Unterkünften von sowohl privaten als auch gewerblichen Vermietern. Mithilfe von Topic Modeling wird eine systematische Übersicht über die Themen akademischer Studien erstellt, die sowohl im zeitlichen Verlauf als auch ihrer Häufigkeit visualisiert werden.

Die Forschenden analysieren dabei 1021 Artikel aus 416 Fachzeitschriften und arbeiten zwei Schwerpunkte innerhalb der untersuchten Artikel heraus: die operationalen Praktiken von AirBnb und die Einflüsse von AirBnb auf andere Bereiche wie etwa urbanen Tourismus oder Immobilien- und Mietpreise. Die zeitliche Analyse zeigt, dass sich die Forschungsschwerpunkte verändert haben: Während anfangs der disruptive Einfluss von Airbnb auf den Immobilienmarkt im Vordergrund stand, fokussieren neuere Studien zunehmend mikroökonomische Aspekte wie Hausordnungen oder das Vertrauen in das Plattformmodell.

In der Studie wird anstelle einer LDA zum Topic Modeling die Methode des Structural Topic Model (STM) von Roberts et al. (2014) verwendet, das es den Forschenden erlaubt, Metadaten in das Modell miteinzubeziehen.

5.3 User review analysis of dating apps based on text mining (Shen et al., 2023)

Shen et al. (2023) analysieren in ihrer Studie negative Userreviews von Dating-Apps mithilfe von Topic Modeling und leiten darauf aufbauend Verbesserungsmaßnahmen ab.

Es werden 142,071 Bewertungen von sechs Dating-Apps ausgewertet. Dabei werden die mithilfe von Topic Modeling die Hauptprobleme der User abgeleitet und in einem zweiten Schritt ein Klassifikationsmodell auf die Sentiment Analysis trainiert und evaluiert. Die Kernproblematiken der App-User sind gemäß der Forschenden Zahlungs- und Abo-Probleme, Fake-Profile und Betrug, Werbung und Abo-Mechanismen und Unzufriedenheit mit dem Matching-System. Die Forschenden schlagen App-Betreibenden vor, dass sie sich dieser Kernproblematiken annehmen durch beispielsweise der Optimierung der Preisgestaltung oder Rückerstattungssysteme, striktere Verifikationsprozesse zur Reduzierung von Fake-Profilen oder die Verbesserung der Matching-Algorithmen, um relevantere Vorschläge zu bieten.

6 Experimentelles Design

Zur Durchführung der Topic Modeling Analyse wurden die Daten in mehreren Schritten aufbereitet und ein passender Corpus generiert. Anschließend wurde ein geeignetes LDA-Modell ausgewählt, nachdem Modelle mit verschiedenen Parametern quantitativ und qualitativ verglichen wurden. Abschließend wurden die extrahierten Themen inhaltlich analysiert.

Die Analyse der Daten erforderte in einem ersten Schritt die Eingrenzung der relevanten Beschreibungstexte. Im Datensatz sind vier Arten von Beschreibungstexten enthalten: Die Beschreibung des Zimmers, der Gegend, des WG-Lebens und Sonstiges. Für den Umfang dieses Projekts wurde nach einer manuellen Prüfung der Beschreibungstexte die Kategorien „WG-Leben“ und „Sonstiges“ ausgewählt, da die übrigen Beschreibungstexte in der Regel nur wenig inhaltliche Aussagen zum WG-Leben, das im Fokus dieser Arbeit steht, enthielten.

Die ausgewählten Beschreibungstexte wurden in einem zweiten Vorverarbeitungsschritt bereinigt. Dazu wurde mithilfe der R-Erweiterung *udpipe* automatisiert ein Lemma-Wörterbuch erstellt mit allen Wörtern des Corpus. Wenige Einträge wurden manuell korrigiert, da sie von dem Tool falsch zugeordnet wurden (zum Beispiel „bist“ $\mapsto$ „bisen“ zu „bist“ $\mapsto$ „sein“). Zur Entfernung von Stop Words wurde eine frei verfügbare Wortliste deutscher Stop Words verwendet (GitHub, 2019). Diese Liste wurde auf den Projektkontext erweitert, wonach häufige Worte wie beispielsweise „Tübingen“ inhaltlich keine informative Relevanz haben. Weiterhin wurden Sonderzeichen aus dem Text entfernt, ebenso wie Begriffe, die nur selten in den Texten auftraten (Nennungen < 5). In einem iterativen Prozess wurden die häufigsten Begriffe des Datensatzes mehrfach geprüft und bei Bedarf entweder in der Lemmatisierung angepasst oder in der Liste der Stop Words hinzugefügt.

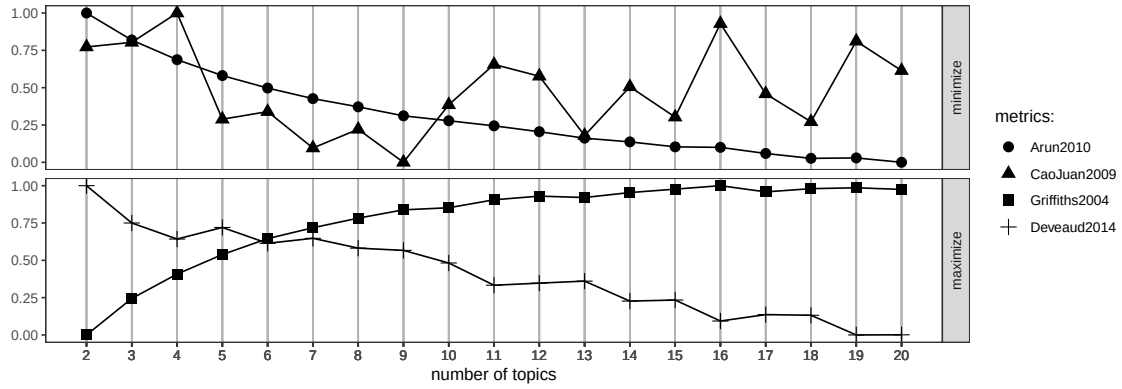

Abbildung 1: Ergebnisse der R-Erweiterung *ldatuning* zur Wahl des Parameters K.

Im folgenden Schritt wurde ein adäquater Parameter K bestimmt, der die Anzahl der Themen des Modells festlegt. Dieser Parameter sollte sowohl quantitativen als auch qualitativen Kriterien genügen. Gemäß der Metriken von Deveaud et al. (2014) und Cao et al. (2009) konnte der optimale Parameter zwischen 5 und 7 eingegrenzt werden (siehe Abb. 1). Nach einer qualitativen Prüfung der so generierten Modelle und ihrer Themen wurde der Parameter auf 5 festgelegt. Die Modelle mit höherer Themenanzahl konnten inhaltlich nicht klar genug differenziert werden.

Die Themenverteilung über den gesamten Datensatz wurde mithilfe der Posteriors der LDA extrahiert. Zur Prüfung der Gruppenunterschiede zwischen FLINTA*- und Non-FLINTA*-WGS wurden die Themenverteilungen für beide Dokumentgruppen separat berechnet.

Im Hinblick auf die Objektivität des experimentellen Designs müssen einige Limitationen beachtet werden. Zum einen erfolgte die Auswahl des finalen Parameters K nach subjektiven Kriterien, sprich, ob einzelne Themen noch abgrenzbar voneinander wahrgenommen wurden. Ebenso ist die Liste der domänenspezifischen Stop Words nach Ermessen der Forschenden ergänzt worden. Noch dazu existiert ein hoher Interpretationsspielraum beim Labeling und der letztendlichen Interpretation der Themen. Schlussendlich kann nicht nur die Interpretation, sondern bereits die Durchführung von Topic Modeling-Methoden verzerrt sein von den Erwartungen und dem Domänenwissen der Forschenden.

7 Ergebnisse und Diskussion

Im Folgenden werden die Ergebnisse der Topic Modeling-Analyse beschrieben. Die Daten sowie die verwendeten Skripte sind online frei verfügbar unter: https://github.com/chill-bird/topic_modeling_wggesucht.

7.1 Themen-Labeling

Nach der Durchführung der LDA ergaben sich fünf Themen, die in Tabelle 1 mit ihrem Anteil im Datensatz und den zehn Worten mit höchster Auftrittswahrscheinlichkeit inner-

No.	Topic Label	Anteil	Top 10 Wörter
1	Alltag, Freizeit &WG-Leben	55%	studieren kochen zeit küche abend nachricht offen wg-leben verbringen kennenlernen
2	Bewerbungsprozess	14%	student mensch person zeit interessen gemeinschaft kennenlernen melden wg-leben verbindung
3	WG-Ausstattung	13%	küche stehen bad student waschmaschinegroß garten keller fahrrad aktuell
4	Nebenkosten	12%	wohnung mieten zeit kurz schön nebenkosten gez strom studieren kosten
5	Englische Anzeige	6%	room thaben buen nieten ther ouen shen abouen gen livehen

Tabelle 1: Topic Distribution für $K = 5$ Topics. Topic Label ist das manuell gesetzte Label. Anteil gibt den Antei des Topics im gesamten Corpus an und ist auf ganze Prozente gerundet. Top 10 Wörter sind die zehn Wörter mit der höchsten Auftrittswahrscheinlichkeit innerhalb eines Topics.

halb eines Themas abgebildet sind. Nach erneuter Prüfung der relevantesten Begriffe in den Themen unter Berücksichtigung der Originaltexte wurden passende Labels vergeben.

Thema 1 *Alltag, Freizeit & WG-Leben* enthielt vor allem Begriffe, die sich auf den Alltag der WG-Bewohnenden inklusive ihrer Freizeitaktivitäten, aber auch das Zusammenleben beziehen könnten. Begriffe wie „kochen“„Freizeit“, „unternehmen“, „Sport“, „Freund“, „treffen“, „Interesse“ unter den 30 Begriffen mit höchster Auftrittswahrscheinlichkeit lassen darauf schließen, dass in diesem Thema vieles über individuelle Hobbys und Interessen in den Beschreibungstexten geschrieben wird. Ebenso präsent sind Begriffe, die sich auf einen gemeinsamen Alltag innerhalb der Wohngemeinschaft beziehen könnten, wie etwa „Putzplan“, „WG-Leben“, „Küche“, „Abend“, „sitzen“, „quatschen“.

Thema 2 *Bewerbungsprozess* enthielt einige Begriffe, die sich auf den Bewerbungsprozess beziehen könnten, sowohl auf die gesuchte Person selbst als auch das Kennenlernen. Dem Bewerbungsprozess zuzuordnen, könnten beispielsweise die Begriffe „Interesse“, „melden“, „Bewerbung“, „kennenlernen“, „Besichtigungstermin“, „vereinbaren“. Die von der WG gesuchte Person könnte über die Begriffe „Student“, „Mensch“, „Person“ gemeint sein.

Thema 3 *WG-Ausstattung* enthielt vornehmlich Begriffe, die die Ausstattung der WG oder auch des gesamten Hauses beschreiben. Einzelne Räumlichkeiten finden sich beispielsweise in den Begriffen „Küche“, „Bad“, „Badezimmer“, „Keller“. Verschiedene Ausstattungsgegenstände werden beispielsweise mit folgenden Begriffen erwähnt: „Waschmaschine“, „Trockner“, „Spülmaschine“, „Kühlschrank“, „Toilette“. Es kommen neben den Substantiven auch beschreibende Attribute wie „groß“ oder „zusätzlich“ unter den wahr-

scheinlichsten Begriffen vor.

Thema 4 *Nebenkosten* enthielt einige Begriffe, die sich vor allem auf zusätzliche Kosten der WG-Anzeige beziehen könnten. Auf die Kosten an sich deuten Begriffe hin wie „Nebenkosten“, „Kosten“, „Euro“. Welche Kosten gemeint sind, ergibt sich aus Begriffen wie „Internet“, „GEZ“, „Strom“, „Wasser“, „Müll“.

Thema 5 *Englische Anzeige* enthält ausschließlich Englische Begriffe, die durch die automatische Englische Lemmatisierung mit *udpipe* häufig falsch zugeordnet wurden (zum Beispiel „that“ ↦ „thaben“). Somit sind die seltsam anmutenden Begriffe zu erklären, die sich in diesem Thema wiederfinden. Für die weitere Analyse wurde dieses Thema vernachlässigt, da die Wörter dieses Themas eindeutig nicht aus der deutschen Zielsprache stammen. Aus den in Abschnitt 4 genannten Gründen konnten jedoch diese Anzeigen nicht entfernt werden.

7.2 Themeninterpretation

Im Nachfolgenden werden die Begriffe, die in den Themen am wahrscheinlichsten vorkommen, interpretiert.

Thema 1 *Alltag, Freizeit & WG-Leben* bildet das Leben der WG-Mitglieder ab. Um einen Eindruck der wahrscheinlichsten Worte zu erhalten, sind diese in Abbildung 2 als Wortwolke abgebildet. Wortwolken der Themen 2-5 können im verlinkten Repository nachvollzogen werden. Bei näherer Betrachtung besonders viele Begriffe, die mit Speisen und Getränken in Verbindung stehen („Kochen“, „Küche“, „Trinken“, „Essen“). Auffällig ist ebenso, dass das Wort „Küche“ das häufigste Wort des gesamten Corpus ist. Die meis-

Abbildung 2: Wortwolke der wahrscheinlichsten Worte von Thema 1 *Alltag, Freizeit & WG-Leben*

ten WGs verbindet, dass die Küche einen Gemeinschaftsraum darstellt, in dem man einander begegnet: ein Treffpunkt, sozialer Mittelpunkt und Ort gemeinsamer Aktivitäten. Kochen und Essen scheinen somit nicht nur alltägliche Notwendigkeiten, sondern auch wichtige soziale Faktoren im WG-Leben zu sein. Das Wort mit der größten Auftrittswahrscheinlichkeit innerhalb von Thema 1 ist der Begriff „studieren“. Dies lässt sich darauf zurückführen, dass viele Beschreibungstexte die einzelnen Mitbewohnenden charakterisiert und vorgestellt werden. Wenn eine Person studiert, so lautet der Satz in vielen der Anzeigen „$< Name >$ ($< Alter >$) studiert ...“. Da Tübingen von überdurchschnittlich vielen Studierenden bewohnt wird („Tübingen“, 2025), ist auch dieser Begriff naheliegend. Generische Überbegriffe für Hobbys sind in diesem Thema ebenso erfasst (beispielsweise „Sport“). Neben allgemeinen Bezeichnungen für Hobbys, etwa „Sport“, lassen sich spezifischere Freizeitbeschäftigungen im Modell weniger deutlich erkennen. Das liegt an ihrer vergleichsweise geringen Worthäufigkeit („Sport“: 101 Nennungen vs. „Volleyball“: 14 Nennungen). Insgesamt zeichnen die relevanten Begriffe des Themas *Alltag, Freizeit & WG-Leben* eine warme, einladende Atmosphäre des Zusammenlebens (beispielsweise „offen“, „Freund“, „entspannt“, „Lust“, „lieben“).

Thema 2 *Bewerbungsprozess* ist geprägt von prozeduralen Begriffen, die sich auf den WG-Bewerbungsprozess über die Plattform *WG-gesucht* beziehen („Zeit“, „Interesse“, „melden“, „anfragen“. „Besichtigungstermin“, „jederzeit“, „beantworten“, „vereinbaren“, „Frage“). Diese Begriffe spiegeln typische Abläufe und Kommunikationsformen innerhalb des Bewerbungsprozesses wider. Auffällig ist zudem, dass die drei wahrscheinlichsten Begriffe innerhalb dieses Themas Personenbezeichnungen sind („Student“, „Mensch“, „Person“). Anhand dieser Begriffe kann vermutet werden, dass entweder die gesuchte Person oder auch die Bewohnenden der aktuellen WG thematisiert werden.

Thema 3 *WG-Ausstattung* befasst sich der wahrscheinlichsten Begriffe nach zu urteilen am meisten mit der Beschreibung der Gemeinschaftsbereiche einer jeden WG und ihrer Ausstattung. Die Begriffe sind in diesem Thema neutral und haben beschreibenden Charakter.

Thema 4 *Nebenkosten* steht ebenso wie Thema 3 in Verbindung mit der Beschreibung der zusätzlichen Kosten. Die wahrscheinlichsten Begriffe umfassende insbesondere die Beschreibung der Nebenkosten, wobei auch die expliziten Begriffe innerhalb des Themas eine hohe Auftrittswahrscheinlichkeit haben („Internet“, „GEZ“, „Strom“, „Wasser“, „Müll“). Darüber hinaus vermischen sich in diesem Thema rein finanzielle Aspekte mit praktischen und organisatorischen Elementen des Zusammenlebens. Begriffe wie „Zusammenleben“, „Zusammenwohnen“, „Putzplan“, „Alltag“, „absprach“ verdeutlichen, dass neben den Kosten auch organisatorische Aspekte der WG eine Rolle spielen.

Wie bereits in Abschnitt 7.1 erläutert, wird von der Interpretation von Thema 5 abgesehen.

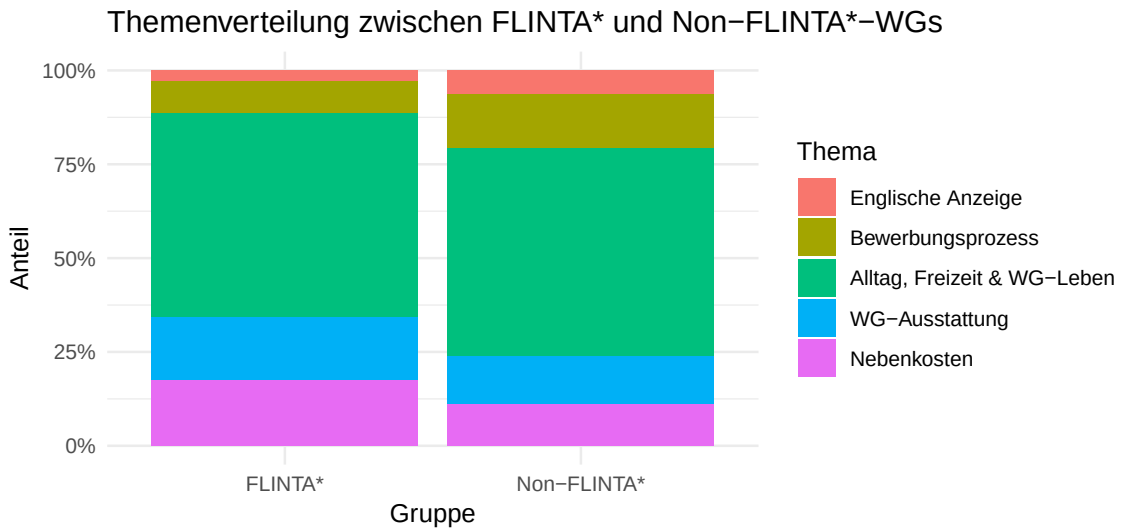

Abbildung 3: Themenverteilung zwischen den Gruppen FLINTA*- and Non-FLINTA*-WG-Anzeigen. Verhältnis von 11.57% FLINTA*-WGs ($n = 61$ Anzeigen) und 88.43% Non-FLINTA*-WGs ($n = 466$ Anzeigen)

7.3 Welche Themen sind in den WG-Beschreibungstexten am stärksten vertreten?

Um die zu Beginn formulierte Forschungsfrage zu beantworten, können nun die Ergebnisse des Topic Modelings herangezogen werden. Zunächst sind thematische Schwerpunkte zu erkennen, die WG-Anzeigen charakterisiert: Die Mitglieder der Wohngemeinschaft, der Bewerbungsprozess, sowie der Wohnraum und die Nebenkosten sind im Topic Modeling erkennbar. Diese thmatische Gliederung spiegelt sich auch in den Platzhaltertexten wider, die *WG-gesucht* beim Erstellen eines Angebots in den Kategorien „WG-Leben“ und „Sonstiges“ vorschlägt: „z.B. Wer sind die Mitbewohner? Wie sieht das Leben in der WG aus? Wer wird gesucht?”“ (Kategorie: WG-Leben), „z.B. Welche Anforderungen muss der Mieter erfüllen? Wie soll bevorzugt Kontakt aufgenommen werden? Zu welcher Uhrzeit?“ (Kategorie: Sonstiges) (WG-Gesucht.de, n. d.). Der stark studentische Charakter der Anzeigen ist in mehreren Themen erkennbar durch einschlägige Begriffe wie „Student“, „studieren“, „Master“ oder „Semester“. Dies könnte auf die Nutzergruppe der Plattform WG-gesucht zurückzuführen sein, aber auch auf den Standort Tübingen als Universitätsstadt („Tübingen“, 2025). Weiterhin kann anhand der Begriffe innerhalb der Themen ein freundschaftliches, offenes Bild des Zusammenlebens in den WG-Beschreibungen festgehalten werden.

7.4 Gibt es Unterschiede zwischen in der Themenverteilung von FLINTA*- und Non-FLINTA*-WGs?

Um der zweiten Forschungsfrage dieser Arbeit nachzugehen, wurden die Themenverteilungen zwischen den beiden Gruppen FLINTA*- und Non-FLINTA*-WGs untersucht. Ab-

bildung 3 zeigt die Themenverteilungen für beide Gruppen. Die genauen Zahlen können in Tabelle 2 nachvollzogen werden. Für die relevanten Themen (1-4) ergeben sich keine nennenswerten Unterschiede zwischen beiden Gruppen. Für das Thema *Alltag, Freizeit & WG-Leben* sind die Verteilungen der beiden Gruppen beinahe identisch. Allerdings lassen sich leichte Tendenzen erkennen: In FLINTA*-WGs nimmt das Thema Bewerbungsprozess einen geringeren Anteil des Beschreibungstextes ein (8.69% vs. 14.25%). Stattdessen wird in FLINTA*-WG-Anzeigen ein größerer Anteil des Textes für die Beschreibung der WG-Ausstattung (16.80% vs. 12.80%) und der Nebenkosten (17.52% vs. 11.02%) verwendet. Angesichts der geringen Unterschiede können jedoch abschließend keine essenziellen Unterschiede zwischen beiden Gruppen festgehalten werden.

No.	**Thema**	**Verteilung Non-FLINTA***	**Verteilung FLINTA***
1	Alltag, Freizeit & WG-Leben	0.5546	0.5414
2	Bewerbungsprozess	0.1425	0.0869
3	WG-Ausstattung	0.1280	0.1680
4	Nebenkosten	0.1102	0.1752
5	Englische Anzeige	0.0647	0.0285

Tabelle 2: Themenverteilung von Non-FLINTA*- and FLINTA*-WG-Anzeigen. Verhältnis von 11.57% FLINTA*-WGs ($n = 61$ Anzeigen) und 88.43% Non-FLINTA*-WGs ($n = 466$ Anzeigen)

7.5 Limitationen

Bei der Interpretation der Ergebnisse sollten einige Limitationen berücksichtigt werden.

Das Differenzieren zwischen dem Alltag der Individuen und dem gemeinsamen WG-Alltag in Thema ist schwierig. Ob die Aktivitäten, die im Thema gelistet sind, tendenziell gemeinsame oder individuelle Unternehmungen sind, kann über die Analyse nicht abschließend geklärt werden und ist vermutlich innerhalb des Themas gemischt. Eine abschließende Aussage über das WG-Leben kann daher nicht getroffen werden. Darüber hinaus wird in den WG-Anzeigen häufig eine eher informelle Sprache verwendet, wobei Begriffe durch ein Lemmatisierungs-Tool potentiell falsch zugeordnet werden könnten. Dazu gehören Neologismen (zum Beispiel „bibben“ für „in der Bibliothek lernen“), Anglizismen (zum Beispiel „viben“) oder auch ungewöhnliche Diminutiva (zum Beispiel „Bierchen“). Auch der Umgang mit Verneinungen stellt in der Analyse eine Herausforderung dar. Ein Beispiel hierfür ist die typische Formulierung „Wir sind keine Zweck-WG!“, von der im Modell lediglich das Wort „Zweck-WG“ übrig bleibt, was die Interpretation erschwert. In ähnlichem Maße sind können trennbare Verben im Deutschen durch Lemmatisierung verfälscht werden (zum Beispiel „Stell dich doch kurz vor“ $\mapsto$ „stellen“, „vor“). Folgt man der Annahme, dass allein Substantive und Adjektive bedeutungsstiftend sind, ist dieser Punkt zu vernachlässigen. Für das vorliegende Projekt ergab sich allerdings bereits aus der geringen Corpusgröße die Notwendigkeit, alle Wortarten in die Analyse miteinzubeziehen.

8 Fazit

Abschließend konnte die vorliegende Projektarbeit vier Themenschwerpunkte identifizieren, die innerhalb der WG-Anzeigen am stärksten vertreten sind. Dazu gehörte der individuelle und gemeinsame Alltag der WG-Mitglieder Eine nähere Aufschlüsselung dieses 1. Themenkomplexes konnte aufgrund der geringen Datenmenge nicht vorgenommen werden, auch wenn einzelne Aspekte (zum Beispiel Essen und Sport) aufschlussreiche Informationen liefern konnten. Es muss festgehalten werden, dass keine eindeutige Trennung zwischen dem Leben der Individuen und dem WG-Leben vorgenommen werden konnte. Klar erkennbar waren ebenso die Themen, die den Bewerbungsprozess sowie räumliche und praktische Gegebenheiten des Zusammenlebens in einer WG beschreiben. Die Platzhaltertexte der Plattform liefern eine mögliche Erklärung dafür, warum insbesondere die ersten beiden Themen in den Anzeigen stark vertreten sind. In diesen wird vorgeschlagen, die WG-Mitglieder und das WG-Leben zu charakterisieren und die Anforderungen an das neue WG-Mitglied zu schildern.

Ein weiterer zentraler Befund ist der stark studentische Charakter der Anzeigen, was sich in häufig verwendeten Begriffen wie „Student“, „studieren“ oder „Semester“ widerspiegelt. Dies könnte sowohl auf die Nutzergruppe von WG-gesucht als auch auf den Standort Tübingen als Universitätsstadt zurückzuführen sein.

Weiterhin ist festzuhalten, dass sich FLINTA*-WG-Anzeigen nicht in hohem Maße von Non-FLINTA*-WG-Anzeigen in ihrer Themenverteilung unterscheiden. Diese Ergebnisse legen nahe, dass sich die sprachlichen und inhaltlichen Schwerpunkte in den Anzeigen beider Gruppen weitgehend ähneln. Somit liefert diese Projektarbeit keinen schlüssigen Erklärungsansatz dafür, warum FLINTA*-WGs in einigen Studien als beliebter wahrgenommen werden. Es ist jedoch denkbar, dass dieser Unterschied nicht in der Struktur der Anzeigen selbst begründet liegt, sondern in anderen Faktoren – etwa der sozialen Dynamik innerhalb der WGs, impliziten Erwartungen an FLINTA*-Wohnformen oder der Wahrnehmung von Sicherheit und Gemeinschaftsgefühl. Diese Aspekte könnten in weiterführenden qualitativen Studien näher untersucht werden.

Insgesamt liefert diese Arbeit eine erste Perspektive auf die Themenstruktur und kommunikative Konventionen von WG-Anzeigen und zeigt gleichzeitig, an welchen Stellen weiterführende qualitative Untersuchungen ansetzen könnten.

Literatur

Blei, D. M. (2012). Topic Modeling and Digital Humanities Journal of Digital Humanities. Verfügbar 28. Februar 2025 unter https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A Density-Based Method for Adaptive LDA Model Selection. *Neurocomputing*, *72*(7), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. (2023). What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures*, *17*, 1–20. https://doi.org/10.1080/19312458.2023.2167965

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Document numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Diehl, C., Andorfer, V. A., Khoudja, Y., & Krause, K. (2013). Not In My Kitchen? Ethnic Discrimination and Discrimination Intentions in Shared Housing among University Students in Germany. *Journal of Ethnic and Migration Studies*, *39*(10), 1679–1697. https://doi.org/10.1080/1369183X.2013.833705

Ding, K., Niu, Y., & Choo, W. C. (2023). The evolution of Airbnb research: A systematic literature review using structural topic modeling. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e17090

GitHub. (2019). Extended List of German Stopwords for Use in Web Projects, Search Engines or Every Thing Else. Verfügbar 28. Februar 2025 unter https://github.com/solariz/german_stopwords/tree/master

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press. Verfügbar 28. Februar 2025 unter https://www.jstor.org/stable/10.5406/j.ctt2jcc3m

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Roth, C. (2019). Digital, digitized, and numerical humanities. *Digital Scholarship in the Humanities*, *34*(3), 616–632. https://doi.org/10.1093/llc/fqy057

Shen, Q., Han, S., Han, Y., & Chen, X. (2023). User review analysis of dating apps based on text mining. *PLOS ONE*, *18*(4), e0283896. https://doi.org/10.1371/journal.pone.0283896

Törnberg, A., & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society*, *27*(4), 401–422. https://doi.org/10.1177/0957926516634546

Tübingen. (2025). *Wikipedia*. Verfügbar 7. März 2025 unter https://de.wikipedia.org/w/index.php?title=T%C3%BCbingen&oldid=253713700
Page Version ID: 253713700.

WG-Gesucht.de. (n. d.). WG Tübingen : WG-Zimmer Angebote in Tübingen WG Tübingen in Tübingen. Verfügbar 7. März 2025 unter https://www.wg-gesucht.de/wg-zimmer-in-Tuebingen.127.0.1.0.html

Woo, A., Cho, G.-H., & Kim, J. (2019). Would You Share Your Home? The Multifaceted Determinants of Preference for Shared Housing among Young Adults. *Applied Geography*, *103*, 12–21. https://doi.org/10.1016/j.apgeog.2018.12.012

Zemanek, .. (1977). Der Index Thomisticus / The Index Thomisticus. *it - Information Technology*, *19*(1-6), 112–122. https://doi.org/10.1524/itit.1977.19.16.112