Martin-Luther-Universität Halle-Wittenberg
Institut für Informatik
Studiengang Informatik (Bachelor)

Expertise Filtering for Social Media Timelines

Bachelorarbeit

Alexander Rensch
geb. am: 18.01.1996 in Hamburg

Matrikelnummer 215233947

1. Gutachter: Prof. Dr. Matthias Hagen
2. Gutachter: Junior-Prof. Dr. Martin Potthast

Datum der Abgabe: 3. Juni 2020

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Halle, 3. Juni 2020

. . . . . .
Alexander Rensch

Zusammenfassung

In dieser Bachelorarbeit wird das Filtern von Tweets auf Basis der Expertise des Autoren untersucht. Hierzu wird der Begriff der Expertise näher definiert und eingegrenzt, sowie der verwendete Celebrity-Korpus ausgewertet. Indem Tweets in einem Vektorraum repräsentiert werden, kann ein Nächster-Zentroid-Klassifizierer trainiert werden, der genutzt wird, um Tweets zu ausgewählten Tätigkeiten im Sinne von Berufen zuzuordnen. Auch wird versucht, eine Support Vector Machine zur allgemeinen Erkennung von Expertise in Tweets zu trainieren. Zu allen vorgestellten Ansätzen werden Experimente durchgeführt und ausgewertet, wobei verschiedene Probleme sichtbar werden.

Inhaltsverzeichnis

1 Einleitung **1**

2 Verwandte Arbeiten **4**

3 Der Begriff der Expertise **6**
3.1 Tätigkeit als Hinweis auf Expertise 7
3.2 Eigenschaften und Kategorien von Tweets 9
3.3 Grenzen des erarbeiteten Begriffs 13

4 Nächster-Zentroid-Klassifizierer **15**
4.1 Repräsentation von Tweets im Vektorraum 16
4.1.1 Datenbereinigung und Normalisierung 17
4.1.2 Feature Extraction 18
4.2 Konstruktion des Klassifizierers 19
4.3 Auswertung 21

5 Weitere Ansätze zur Klassifizierung **23**
5.1 Zentroiden aus Wikipedia-Artikeln 24
5.2 SVM zur Bestimmung von Expertise-Tweets 26

6 Fazit und zukünftige Arbeit **28**

Literaturverzeichnis **30**

Abbildungsverzeichnis

1.1 Jürgen Klopp zu Expertenmeinungen in der Corona-Krise 2

3.1 Tätigkeiten in Wikidata 8
3.2 Beispiel: Nicht entscheidbarer Tweet 10
3.3 Beispiele der Kategorien 11

4.1 Pipeline des Expertise-Filters 15
4.2 Schritte zur Bereinigung der Tweets 17
4.3 Vektorrepräsentation von Dokumenten wie in [11] 19
4.4 Auswertung: Nächster-Zentroid-Klassifizierer 22

5.1 Auswertung: Nächster-Zentroid-Klassifizierer (Wikipedia) 25
5.2 Auswertung: Support Vector Machine 27

Kapitel 1

Einleitung

In dieser Arbeit soll das Filtern von kurzen Texten aus sozialen Medien, hier in Form von Tweets der bekannten Social Media Plattform Twitter, auf Basis der Expertise ihrer Autoren untersucht werden. Ziel ist es, nur diejenigen Tweets zu erhalten, die der Expertise des Autoren entsprechen, also zum Beispiel ein Tweet eines Informatikers über die neuesten Entwicklungen im Bereich des maschinellen Lernens.

Die Timelines in den sozialen Medien sind überflutet mit privaten Statusmeldungen, Neuigkeiten, Angeboten zu Deals, Kommentaren und Meinungsbekundungen, sowie anderen geteilten Inhalten, die gleichwertig nebeneinander stehen und weder kategorisiert noch anderweitig gekennzeichnet sind. Je mehr Nutzern man dort „folgt“, d.h. ihre veröffentlichen Inhalte abonniert, desto schwerwiegender wird diese Problematik. Dabei gibt es viele unterschiedliche Personengruppen, denen es lohnt zu folgen. Private Kontakte aus dem Familien- oder Freundeskreis, Bekannte, Arbeitskollegen oder andere interessante, bekannte Persönlichkeiten liefern einem täglich Neuigkeiten, die aber nicht auf ein Themenfeld beschränkt sein müssen. Sogar Unternehmen oder Organisationen sind in den sozialen Medien vertreten, die ihrerseits zum Beispiel mit spannenden Ankündigungen oder aber mit Werbung aufwarten. Diese Masse an Informationen erschwert es erheblich, sich schnell einen Überblick über aktuelle Meldungen zu bestimmten Themenfeldern zu verschaffen, die einen derzeit interessieren.

Weiter äußern sich zu den besagten Themen auch Autoren, die nicht notwendigerweise Expertise in diesen Gebieten besitzen. Die Relevanz dieser Problemstellung zeigt sich aktuell zum Beispiel anhand der Corona-Pandemie. Menschen sind mit einer Vielzahl an (teils auch gegensätzlichen) Aussagen konfrontiert, die eingeordnet werden müssen, um das eigene Verhalten bestimmen zu können. Diese Aussagen werden nicht nur von Experten getroffen, sondern auch grundsätzlich von (berühmten) (Privat-)Personen mit Reichweite in

den sozialen Netzen, was neben Falschinformation und Verwirrung auch Expertenmeinungen verschleiern kann. Jürgen Klopp, der sehr bekannte Manager vom FC Liverpool, äußerte hierzu in einem Interview auf Nachfrage, was er zu der Corona-Pandemie zu sagen hätte, dass nur „Menschen mit Fachwissen“ über das neue Coronavirus sprechen sollten und seine eigene Meinung „nicht wichtig“ sei. Während das Filtern von Aussagen und Meinungen in einem öffentlichen Raum, wie die sozialen Medien inzwischen welche geworden sind, stets kritisch zu betrachten ist, so ist die generelle Einschätzung, ob eine Aussage eines Autoren zu seiner Expertise zählt, eine wertvolle Zusatzinformation zur Einordnung seines Beitrags innerhalb der Diskussion.

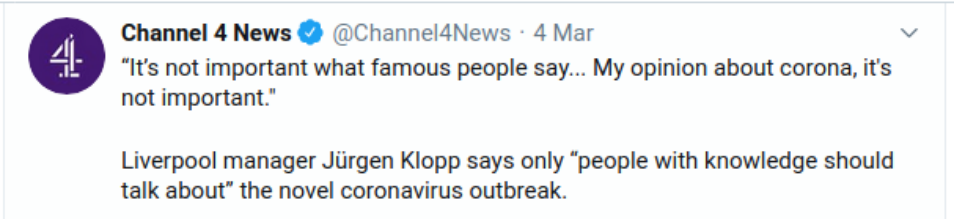


Abbildung 1.1: Jürgen Klopp zu Expertenmeinungen in der Corona-Krise

Quelle: `https://twitter.com/Channel4News/status/1235178131880841216`

Um einen Expertise-Filter zu konstruieren, muss der Begriff der Expertise genauer definiert werden. Als Ansatzpunkt wird die Tätigkeit einer Person herangezogen, die als Hinweis auf ihre Expertise dienen soll. Hierfür wird ein vorgegebener Datensatz aus Tweets von prominenten, englischsprachigen Twitter-Nutzern verwendet, der zusätzlich Informationen zu diesen Tätigkeiten der Autoren (engl. „occupations“) bietet und somit die Themenfelder anzeigt, in denen ein Autor Expertise aufweist. Ein Autor, der laut Datensatz als Informatiker beschäftigt ist, besitzt demnach Expertise in der Informatik, jedoch keine in wirtschaftlichen oder geisteswissenschaftlichen Fragen. Es wird insgesamt eine Auswahl von 20 Tätigkeiten getroffen, die in dieser Arbeit näher untersucht werden. Die verbleibenden Tweets werden in Hinblick auf ihre besonderen Eigenschaften vorverarbeitet. Die tatsächliche Expertise einer Person muss sich in Wahrheit nicht nur auf das Feld ihrer Tätigkeit erstrecken. Jedoch soll hiermit ein erster, sicherer Ansatz geschaffen werden, mit dem für jedes Profil mindestens ein Expertise-Feld zur Verfügung steht, womit die untersuchten Tweets verglichen werden können.

Tweets werden in einem hochdimensionalen Vektorraum repräsentiert, wobei dem Bag-of-Words-Ansatz gefolgt wird. Schließlich sollen Tweets den durch die Tätigkeiten vorgegebenen Themenfeldern mittels eines Nächster-Zentroid-Klassifizierers zugeordnet werden, indem für jede Tätigkeit ein Zentroid auf Basis der Tweet-Vektoren gebildet wird, deren Autoren diese Tätigkeit ausüben.

Nach der Klassifikation entscheidet der Expertise-Filter anhand der gegebenen Expertise des Autoren, ob der Tweet erhalten oder ausgeblendet werden soll. Die Auswertung des Klassifizierers zeigt allerdings, dass dieser stark überfiltert, also zu viele Expertise-Tweets entfernen würde, und insgesamt nur eine geringe Treffergenauigkeit aufweist. Dies ist besonders darauf zurückzuführen, dass ein wesentlicher Anteil an Tweets, auf die der Klassifizierer trainiert wird, keine Expertise aufweist und somit eine Vielzahl ungeeigneter Tweets in die jeweiligen Zentroiden einfließen.

Zusätzlich zur Konstruktion eines Nächsten-Zentroid-Klassifizierers auf Basis von Tweets werden noch zwei weitere Ansätze vorgestellt. Zum einen wird der ursprüngliche Klassifizierer nicht mehr auf Tweets trainiert, sondern stattdessen auf Wikipedia-Artikeln zu den Tätigkeiten bzw. den Tätigkeitsfeldern. Dies geschieht aus der Überlegung heraus, möglichst viel spezifisches Fachvokabular in die Zentroiden einzubetten und dem Problem zu begegnen, dass die vorher gebildeten Zentroiden zu viele „Nicht-Expertise-Tweets“ enthielten. Zum anderen wird eine One Class Support Vector Machine auf Basis von 1000 per Hand ausgewerteten Tweets erstellt, die die Klasse von „Expertise-Tweets“ im Allgemeinen bestimmen soll. So soll unter anderem eine Vorauswahl der Tweets getroffen werden können, die weiter untersucht und zur Modellbildung herangezogen werden,Keiner der beiden Ansätze zeigt vielversprechende Ergebnisse. Zukünftige Arbeit sollte besonders die Bereinigung und Normalisierung von Tweets in den Vordergrund stellen. Außerdem weisen Tweets einige generelle, nicht vom Thema abhängige Eigenschaften auf, die für weitere Experimente zur Einteilung von besprochenen Kategorien betrachtet werden können.

Kapitel 2

Verwandte Arbeiten

Einen Filter auf Basis von Expertise der Autoren in den sozialen Medien zu konstruieren, ist eine neue und bisher offenbar unerforschte Aufgabenstellung, zu der wenig bis keine Literatur existiert. Sie umfasst jedoch im Wesentlichen zwei größere Themenfelder, zu denen bereits einige Autoren gearbeitet haben und deren Lösungen kombiniert werden müssen. Zum einen muss die Expertise der Autoren ermittelt werden. Dies umfasst das Forschungsgebiet des Expertise Retrievals, spezieller des Expert Profilings, das versucht, die Frage zu beantworten, in welchen Themen eine Person ein Experte ist. Zum anderen müssen die von den Nutzern verfassten Kurztexte den Themen zugeordnet werden können, die im Rahmen des Profilings definiert wurden. Dies umfasst das weite Feld der Textklassifizierung, speziell der Kurztextklassifizierung, das die Forschung bis heute vor Probleme stellt.

Die Herausforderungen und möglichen Ansätze zum Expertise Retrieval werden in [1, 2] betrachtet. Hinweise auf Expertise werden hierbei im Wesentlichen in den von den Autoren verfassten Dokumenten gesucht, wobei auch andere Quellen zur Bestimmung von Expertise genannt werden. So können Informationen aus dem sozialen Umfeld der Personen erworben werden und - im beruflichen Kontext - von Arbeitskollegen oder dem Unternehmen an sich. In [12, 17] wird gezeigt, dass gerade der Fokus auf die rein textuelle Ebene zur Bestimmung von Expertise, zumindest auf Twitter, nicht zielführend ist, sondern die Zugehörigkeit zu sogenannten „Listen“, in denen Nutzer andere Nutzer als relevante Quellen für Themengebiete zuordnen, besonders wertvoll ist. In [10] wird zudem das Problem angesprochen, dass bisherige Methoden des Expertise Retrieval nicht auf Änderungen von Expertise im Laufe der Zeit eingehen. Hierfür entwerfen sie hierarchische Expertise-Profile, die die Expertise-Felder in einer Taxonomie anordnen und mit einem Zeitstempel versehen sind, um genannte Änderungen nachvollziehen zu können.

Einen umfangreichen Datensatz zu konstruieren, der Daten, speziell Texte,

aus den sozialen Medien enthält und zugleich Informationen über die Expertise ihrer Autoren, ist selbst eine komplexe Problemstellung. In [15] wird der „Webis Celebrity Corpus 2019“ (kurz: Celebrity-Korpus) vorgestellt, der zu einer Vielzahl von englischsprachigen, prominenten Twitter-Nutzern Tweets und Profilinformationen vereint. Darin enthalten sind unter anderem die Tätigkeiten (engl. „occupations“) der erfassten Twitter-Nutzer. Im Rahmen des „PAN Celebrity Profiling Tasks“ sollen nur mithilfe der Tweets die Eigenschaften der Autoren „Gender“, „Alter“ und „Tätigkeit“ vorausgesagt werden. Der Celebrity-Korpus wird in dieser Arbeit als Datengrundlage dienen, sodass sich die Experimente auf die Plattform Twitter beschränken. Wie in Kapitel 3 ausführlich erläutert, wird die im Korpus angegebene Tätigkeit als Referenz für den hier verwendeten Expertise-Begriff dienen.

Textklassifizierung ist ein wichtiger Teil des Text Data Minings und verwendet Techniken des maschinellen Lernens. In [16] werden der Textklassifizierungsprozess allgemein und gängige Text-Klassifizierer vorgestellt, darunter der Nächster-Zentroid-Klassifizierer und die Support Vector Machine, die in dieser Arbeit zum Einsatz kommen. In [11] wird das zugrundeliegende Vektorraummodell erläutert und in [8, 13] das verwendete TFIDF-Gewichtungsmodell bzw. dessen konkrete Implementierung mittels *Scikit-learn* beschrieben. In [4] wird das Konzept eines Nächster-Zentroid-Klassifizierers näher erläutert und analysiert, während in [7] die OneClass-SVM zur Verwendung in der Dokumentklassifikation vorgestellt wird. Die OneClass-SVM fällt in den Bereich des nicht-überwachten Lernens, die zum Beispiel zur *Novelty Detection* verwendet wird. Hier soll „Expertise“ die angelernte Klasse sein und Tweets darauf überprüft werden, ob sie eine „Neuheit“ darstellen.

Die speziellen Eigenschaften von Kurztexten, wie sie auf Plattformen sozialer Medien verfasst werden, stellen weitere Herausforderungen dar. In [14] wird ein Ansatz zur Kurztextklassifikation auf Twitter vorgestellt, der Tweets in zuvor definierte Kategorien wie Neuigkeiten, Meinungen, Deals, Events oder private Nachrichten einteilt. Die theoretische Arbeit zur Beschaffenheit der Tweets auf Twitter wird auch hier Beachtung finden. In [3, 9] wird versucht, Tweets nach ihrem Thema zu kategorisieren. Zusammen mit [5] werden die besonderen Eigenschaften von Tweets und die damit verbundenen Probleme analysiert und Verfahren vorgestellt, um Tweets zuverlässig vorzuverarbeiten. Unter anderem wird das „Lexical Normalization“ eingeführt, das versucht, vokabularfremde Wörter (OOV-Wörter) auf Wörter zurückzuführen, die im Vokabular vorhanden sind (IV-Wörter).

Kapitel 3

Der Begriff der Expertise

Die zentrale Frage, die der Expertise-Filter beantworten muss, ist, ob ein Tweet zur Expertise seines Autoren gehört oder nicht. Expertise lässt sich beschreiben als breite und ausgeprägte Kompetenz bzgl. Wissen und Erfahrung in einem bestimmten Themenfeld, die eine Person folglich als Experten dafür auszeichnet. Somit können Personen Expertise in mehreren, verschiedenen Themenfeldern aufweisen, wobei in dieser Arbeit nur Profile mit einfacher Expertise betrachtet werden.

Die Expertise von Personen zu bestimmen, ist ein komplexes Problem, das besonders auch im unternehmerischen Kontext bearbeitet wird. So wird versucht, im Unternehmen oder einer Organisation einen Experten zu einem bestimmten Themenfeld zu finden oder zu einer Person alle Themenfelder zu ermitteln, in denen diese Experte ist. Dies dient beispielsweise dazu, schnell einen sachkundigen Kollegen ausfindig zu machen, der eine relevante Fragestellung am besten bearbeiten kann. Diese zwei Aufgaben, das *Expert Finding* und *Expert Profiling*, werden dem *Expertise Retrieval* zugeordnet und können „als zwei Seiten einer Medaille“ betrachtet werden. Für beide Aufgaben muss die zentrale Frage beantwortet werden, ob eine Person zu einem untersuchten Thema Expertise aufweist oder nicht. Dabei werden zu den betrachteten Personen Sachprofile erstellt, die ihre Kompetenz zu verschiedenen Themenfeldern in Form eines Punktestands erfassen. [1, 2, 10] Eine solche feingliedrige Unterschiedung im Grad der Expertise wird in dieser Arbeit nicht vorgenommen. Hier werden vereinfacht nur die beiden Fälle betrachtet, ob eine Person (einen beliebigen Grad an) Expertise aufweist oder nicht. Somit kann die Expertise einer Person als einfache Liste der Themenfelder repräsentiert werden, in denen sie Expertise aufweist.

Um eine geeignete Datengrundlage zu schaffen, die die wahre Expertise von Autoren offenlegt, müssten Ansätze des Expertise Profiling verfolgt werden oder manuell die Expertise von untersuchten Twitter-Profilen bestimmt

werden. Während eine manuelle Annotation mit sehr großer Mühe und nicht überschaubaren Zeitaufwand verbunden wäre, stellt die Beschaffenheit der Kurztexte aus Twitter die Ansätze des Expertise Profiling vor schwerwiegende Probleme. Experimente in [17] zeigen, dass die vielversprechendsten Hinweise auf Expertise gar nicht in den Tweets selbst zu finden sind, sondern in Meta-Informationen zu den Twitter-Profilen, ganz besonders der Zugehörigkeit zu sogenannten *Listen*. Diese Listen werden von Nutzern erstellt und tragen meist prägnante Namen von Themen, zu denen die zugeordneten Nutzer als relevante Informationsquelle dienen. Eine solche Liste könnte zum Beispiel den Namen „Machine Learning“ tragen und würde Profile von Informatikern aus diesem Bereich führen. Um dieser Problematik auszuweichen, wird daher stattdessen in dieser Arbeit auf die *Tätigkeit* als Hinweis auf Expertise zurückgegriffen und ein Datensatz vorgestellt, der umfangreiche Twitter-Profile mit Informationen zu den Tätigkeiten ihrer Autoren besitzt.

Eine zusätzliche Schwierigkeit bei der Konstruktion des Expertise-Filters ist seine Anwendung auf Kurztexte aus den sozialen Medien, speziell die Tweets von Twitter. In gängigen Ansätzen des Text Data Minings wird mit längeren Dokumenten gearbeitet, die eine Vielzahl an Worten enthalten, wodurch ihr Inhalt zuverlässig nur durch Betrachtung der einzelnen Worte bestimmt werden kann. Die hier untersuchten Tweets weisen jedoch neben ihrer Kürze noch weitere Eigenschaften auf, die eine Textanalyse erschweren und genauer untersucht werden. Bei genauerer Betrachtung können zudem verschiedene Kategorien gebildet werden, in die sich die Tweets einordnen lassen.

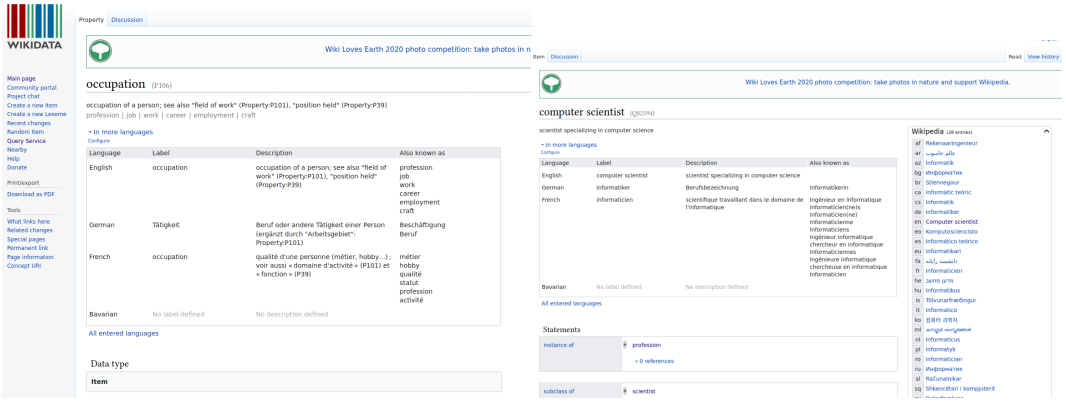
Die hier vorgestellte Sichtweise auf Expertise und deren Annahmen weisen einige Limitierungen auf. Diese werden schließlich besprochen und kurz weitere Problemfelder thematisiert.

3.1 Tätigkeit als Hinweis auf Expertise

Um Experten zu finden, wird gewöhnlich versucht, Beziehungen zwischen Personen und Themenfeldern aufzudecken. Als Hinweis auf Expertise wird häufig das gleichzeitige Auftreten des Namens der untersuchten Person mit den relevanten Expertise-Themenfeldern betrachtet. [1] Weitere Hinweise sind insbesondere auch im unternehmerischen Umfeld der Person zu finden. Teile ihrer Kompetenz und Fähigkeiten können von Arbeitskollegen oder der Organisation an sich abgeleitet werden. [2] Infolgedessen wird in dieser Arbeit die (berufliche) Tätigkeit einer Person als Anhaltspunkt für ihre Expertise verwendet.

Korpora, die eine große Anzahl an Tweets von vielen Profilen beinhalten und gleichzeitig Informationen über diese Profile anbieten, sind schwierig zu konstruieren. Als Datengrundlage wird hier der „Webis Celebrity Corpus 2019“

(kurz: Celebrity-Korpus) aus [15], der eine Vielzahl an Tweets von englischsprachigen Prominenten (engl. „Celebrities“) enthält sowie eine Liste von Eigenschaften, darunter den Tätigkeiten (engl. „occupations“) dieser Prominenten, die ein Profil der Nutzer zeichnen. Tätigkeiten umfassen beispielsweise „Politiker“, „Reporter“ oder „Mathematiker“ und offenbaren somit das Berufsfeld, in denen diese Personen tätig sind. Dieser Korpus wurde konstruiert, indem 71706 Twitter-Konten mit ihren Wikidata-Items verknüpft wurden. Als Prominenter wird bezeichnet, wer sowohl ein verifiziertes Twitter-Konto besitzt als auch *bedeutend* genug ist, dass er oder sie Thema eines Wikipedia-Artikels ist und einen Wikidata-Eintrag besitzt. Jedes Profil enthält im Durchschnitt 2181 Tweets, sodass der Korpus insgesamt 156411899 Tweets umfasst. Durchschnittlich 989 Tweets eines jeden Profils sind länger als 20 Zeichen und 77 Prozent aller Timelines bestehen aus nur englischsprachigen Tweets. [15]



(a) Item: Tätigkeit **(b)** Beispiel: Informatiker

Abbildung 3.1: Tätigkeiten in Wikidata

Quelle: `https://www.wikidata.org/`

Die in Wikidata angegebenen Tätigkeiten dienen als Grundlage für unser Verständnis von Expertise. Wikidata ist ein Speicherplatz für strukturierte Daten, wobei Wissen in Form von Objekten, die mit anderen Objekten verknüpft sind, angegeben wird. In Abbildung 3.1a ist der Wikidata-Eintrag zum Eigenschafts-Objekt „occupation“ dargestellt, das die Oberklasse aller betrachteten Tätigkeiten bildet. Tätigkeiten können ebenfalls in einem Klassenverhältnis zueinander stehen. So ist der in Abbildung 3.1b angegebene „Computer Scientist“ zum Beispiel eine Unterklasse des „Scientist“. Hieraus lässt sich eine Taxonomie der Tätigkeiten ableiten, die es ermöglichst, in der Struktur tiefer angesiedelte Tätigkeiten mit ihren Eltern-Tätigkeiten zu assoziieren, um so inhaltlich ähnliche Objekte zusammenzufassen und besser abgegrenzte Klassen zu entwerfen.

In dieser Arbeit wird die Annahme getroffen, dass ein Nutzer, der in einem bestimmten Feld tätig ist, auch darin Expertise aufweist. Ein als Informatiker tätiger Nutzer wird also kompetente Aussagen zur Informatik treffen können, wie zum Beispiel in Bezug auf aktuelle Entwicklungen im Bereich des maschinellen Lernens. Da im Voraus nicht bekannt ist, welche Tweets eines Profils im Speziellen auch das Themenfeld der Tätigkeit behandeln, wird weiter angenommen, dass *alle* Tweets der genannten Tätigkeit zuzuordnen sind. Um zu vermeiden, dass Tweets von Profilen stammen, die in mehreren Tätigkeitsfeldern beschäftigt sind und somit nicht eindeutig einem Tätigkeitsfeld zugeordnet werden können, werden nur Profile betrachtet, die genau eine Tätigkeit in Wikidata angegeben haben.

Die Annahme, dass alle Tweets eines Nutzers auf Twitter immer in der Funktion als Professioneller bzw. Experte geschrieben werden, ist stark vereinfachend und in den meisten Fällen auch falsch. Twitter als soziales Netzwerk deckt offensichtlich auch persönliche Nachrichten und Meinungen zu allen Bereichen ab - auch zu denen, für die die Nutzer keine geprüfte Kompetenz aufweisen. Dies ist ein Problem, da sich so zu jeder Tätigkeit auch Tweets sammeln, die eben keine Expertise aufweisen und auch keinen Anspruch daran haben. Weiter sind diese Kategorien von Tweets grundsätzlich nicht tätigkeitsspezifisch, was bedeutet, dass jede Tätigkeit folglich ähnliche Tweets enthalten kann, was ihre Unterscheidung noch zusätzlich erschwert. Jeder Nutzer kann Glückwünsche empfangen oder versenden, Situationen aus Familie und Freizeit kommentieren oder einfach seine Gedanken aus dem Alltag teilen, die in Hinblick auf Expertise jedoch nicht relevant sind. Diese Einschätzung wird in [17] gestützt, deren Experimente implizieren, dass „eine große Varianz von Informationen“ in Tweets und Retweets existiert und Experten eines bestimmten Themenfeldes nicht notwendigerweise die gesamte Zeit oder überhaupt über dieses Thema schreiben.

Die geschilderte Problematik wird bei der Konstruktion des Expertise-Filters in dieser Arbeit vernachlässigt und soll Teil zukünftiger Arbeit sein. Es erscheint jedoch vielversprechend, im Voraus Kategorien von Tweets festzulegen, die überhaupt auf Expertise untersucht und schließlich auch zur Modellbildung für einen Klassifizierer herangezogen werden sollen.

3.2 Eigenschaften und Kategorien von Tweets

Um ein besseres Verständnis für die Beschaffenheit der untersuchten Tweets im Celebrity-Korpus und eine Grundlage zur Bewertung des Expertise-Filters zu bekommen, wurden insgesamt 1200 Tweets per Hand ausgewertet. Diese wurden auf einer dreistufigen Skala eingeordnet, wobei zwischen folgenden Stufen

unterschieden wurde:

- (1) Expertise-Tweets, die eindeutig der Expertise des Autoren entsprechen
- (0) Nicht entscheidbare Tweets
- (-1) Nicht-Expertise-Tweets, die eindeutig nicht der Expertise des Autoren entsprechen oder keine Aussagen bzw. Meinungen mit Anspruch auf Expertise kundtun.

Von den 1200 untersuchten Tweets wurden 460 als Expertise-Tweets identifiziert, die der Expertise des Autoren entsprechen und somit weiterhin sichtbar bleiben sollten. 604 Tweets wurden als Nicht-Expertise-Tweets erkannt, die entweder Aussagen zu Themenfeldern trafen, die nicht in das Feld der Expertise des Autoren fallen, oder grundsätzlich keinen Anspruch auf Expertise haben. Für die restlichen 136 Tweets war die Unterscheidung nicht eindeutig möglich, da die Aussage entweder für den Annotator aufgrund mangelnder Kenntnis im Themenfeld nicht genau zugeordnet werden konnte, dem relevanten Themenfeld grundsätzlich nur entfernt zugeordnet werden konnte oder zwar inhaltlich dem Expertise-Themenfeld des Autoren zuordnen ließ, jedoch nicht sachlich war oder nicht für die Allgemeinheit relevant. Manche Tweets enthalten zum Beispiel humoristische oder polemische Elemente, die den Sinngehalt der Aussage verschleiern können, sowie (versteckte) Werbung. Ein Beispiel für einen solchen nicht eindeutigen Tweet ist in Abbildung 3.2 zu sehen, der von einem Informatiker retweetet wurde. Der geteilte Artikel behandelt ein Vorhaben zur Digitalisierung in der Bildung, ist jedoch nicht eindeutig der Informatik zuzuordnen.

Abbildung 3.2: Beispiel: Nicht entscheidbarer Tweet

Quelle: `https://twitter.com/jandersonQZ/status/955400385602211840`

Der erhöhte Anteil von nicht entscheidbaren Tweets zeigt die Schwierigkeit des zugrundeliegenden Problems auf, einen Expertise-Filter zu konstruieren. Mehr als jeder zehnte Tweet konnte nicht eindeutig in Bezug auf seine Expertise beurteilt werden. Bereits für Menschen ist es schwierig, den Sinngehalt eines kurzen Textes in Form eines Tweets zu erfassen und im Kontext der relevanten Tätigkeit des Autoren richtig einzuordnen. Auch die Entscheidung, welche Aussage tatsächlich als sachliche „Expertenmeinung“ gewertet werden soll und welche nur in der Rolle als Privatperson getätigt wurde, kann nicht immer eindeutig und zufriedenstellend getroffen werden. Hier muss die Frage geklärt werden, wie diese Grenzfälle behandelt werden. Sollen nicht eindeutig klassifizierbare Tweets dennoch entfernt werden oder sollen sie beibehalten werden, um das Risiko zu minimieren, wertvolle Informationen zu verlieren? Es erscheint plausibel, dass der zu entwerfende Filter eine konservative Strategie nutzen sollte, die im Zweifel auch diese Tweets erhält. Daher werden die nicht entscheidbaren Tweets bei späterer Verwendung auch zu den Expertise-Tweets gezählt.

Aus der manuellen Auswertung geht weiter hervor, dass ungefähr die Hälfte der veröffentlichten Tweets gar keine Expertise enthält. Dieses Ergebnis bekräftigt die zuvor geschilderte Problematik in Bezug auf die Annahme, dass die Themenfelder der Tätigkeiten tatsächlich durch alle Tweets von Autoren, die der jeweiligen Tätigkeit zugeordnet sind, beschrieben werden können. Ungefähr jeder zweite Tweet verzerrt die Repräsentation der Tätigkeit.

Die manuelle Durchsicht ließ Kategorien erkennen, in die sich Tweets im Allgemeinen einteilen lassen. Unterschieden werden kann zwischen (neutralen, persönlichen und meinungsgebundenen) Neuigkeiten, Events, Meinungen, Deals und privaten Nachrichten wie es auch in [14] geschieht.

(a) Beispiel: Neutrale Neuigkeiten **(b)** Beispiel: Private Nachrichten

Abbildung 3.3: Beispiele der Kategorien

Neutrale Neuigkeiten können von einem breiteren Publikum verstanden werden und präsentieren nur Fakten, keine Meinungen. Normalerweise sind sie sehr strukturiert und enthalten keine Rechtschreibfehler oder andere Auffälligkeiten. Neben einer groben Zusammenfassung der Neuigkeit wird auch ein Link zu einem detaillierteren Artikel beigefügt.

Persönliche Neuigkeiten behandeln gewöhnlich den Gedankengang eines einzelnen Nutzers oder eine Beschreibung seiner aktuellen Situation. Sie sind meist nur für einen kleinen Kreis von Nutzern bedeutend. Wie neutrale Neuigkeiten enthalten sie keine Meinungen, sondern informieren direkt, aber ohne weitere Links. Sie sind nicht notwendigerweise strukturiert und können auch Wortverkürzungen enthalten, um bspw. die Tweet-Längenbegrenzung einzuhalten.

Meinungsgebundene Neuigkeiten beschreiben eine positive oder negative Meinung, die durch das Thema der Neuigkeit selbst ausgedrückt wird. Es handelt sich hierbei beispielsweise um veröffentlichte Kommentare und müssen nicht zwangsläufig die Meinungs des Tweet-Autoren selbst widerspiegeln. Somit sind sie auch von persönlichen Meinungen abzugrenzen.

Meinungen sind ähnlich zu meinungsgebundenen Neuigkeiten, behandeln jedoch immer die Meinung des Autoren in Bezug auf bestimmte Themen. Sie können Abkürzungen enthalten und entweder durch unnatürliche Kapitalisierung oder Buchstabenwiederholungen betont werden. Zudem verwenden sie Emoticons.

Deals behandeln Angebote zu Produkten oder Dienstleistungen. Sie behandeln häufig Links zu detaillierteren Beschreibungen und sind grundsätzlich strukturiert und enthalten keine Rechtschreibfehler oder Abkürzungen. Zudem ist das Vokabular auf einschlägige Begriffe wie „Angebot“ oder „kostenlos“ begrenzt. Teilweise befinden sie sich an der Grenze zu Spam.

Events umfassen Tweets zu bestimmten Veranstaltungen. Sie enthalten Informationen zu Teilnehmern, dem Ort des Geschehens und zur Zeit. Aufgrund der Zeichenbegrenzung auf Twitter sind sie nicht immer vollständig strukturiert und können auch nur eine Teilmenge der genannten Informationen enthalten.

Private Nachrichten sind nur an bestimmte Nutzer auf Twitter addressiert. Sie sind charaktierisiert durch Erwähnungen anderer Nutzer und nicht interessant oder verständlich für Außenstehende. Sie können Rechtschreibfehler und Abkürzungen enthalten. Inhaltlich können sie zudem Meinungen oder Informationen zu Events enthalten.

Die vorgestellten Kategorien lassen einige strukturelle und stilistische Merkmale erkennen, anhand derer sich Tweets voneinander unterscheiden lassen, ohne ihren konkreten Inhalt bzw. ihre Aussage zu untersuchen. Sie lassen sich einfach, zum Beispiel durch reguläre Ausdrücke, aus den Tweets extrahieren. Wichtige Kennzeichen sind das Vorhandensein von:

- Erwähnungen (Bsp.: „@user“)
- Retweet-Symbol („RT“)
- Links
- Hashtags
- Emoticons
- Zeit-Informationen
- Währungs- oder Statistikinformationen
- Slang oder Abkürzungen (Bsp.: „lol“ für „laughing out loud“)
- Rechtschreibfehler
- Wortbetonungen durch Kapitalisierung oder Buchstabendopplungen

Slang bzw. Abkürzungen, Rechtschreibfehler und Wortbetonungen lassen sich zusammenfassen als das Auftreten von vokabularsfremden Wörtern (OOV-Wörter, engl. „out-of-vocabulary“) wie es in [5] näher beleuchtet wird. Weniger als die Hälfte der Nachrichten auf Twitter enthalten OOV-Wörter. Ungefähr 15% der Tweets auf Twitter weisen 50% oder mehr OOV-Wörter auf, wobei es sich größtenteils um verformte, ungrammatikalische Wörter und Eigennamen handelt. Bei den verformten Wörtern handelt es größtenteils um Buchstabenfehler, bei denen Buchstaben entweder fehlen oder zusätzliche Buchstaben vorhanden sind. Aber auch Slang ist mit mehr als 10% der Fälle vertreten.

Dieser große Anteil an Unregelmäßigkeiten stellt zusätzliche Herausforderungen in der Textanalyse dar, die in Kapitel 4.1.1 näher besprochen werden. Die genannten Merkmale legen nahe, dass sich Tweets auch ohne Bezug auf ihren thematischen Inhalt untersuchen lassen. Eine Vorklassifizierung in Bezug auf die vorgestellten Tweet-Kategorien zur Auswahl einer geeigneten Teilmenge von Tweets bietet sich an, um dem zuvor geschilderten Problem zu begegnen, dass den Tätigkeiten zu viele nicht relevante Tweets zugeordnet sind.

3.3 Grenzen des erarbeiteten Begriffs

Der hier vorgestellte Begriff von Expertise schreibt einer Person bisher nur ein einziges Themenfeld zu, zu dem sie Expertise aufweisen kann. Während die Erweiterung auf mehrere Themenfelder durch die Hinzunahme weiterer Tätigkeiten einer Person nur in Bezug auf die Bestimmung ihrer Expertise unproblematisch ist, würde es die Zuordnung der Tweets der Autoren zu den

Tätigkeiten mit jeder neuen Tätigkeit erheblich erschweren. Für eine als Autor und zugleich Musiker gekennzeichnete Person müsste entschieden werden, ob und welche Tweets in die Tätigkeiten einfließen. Dies führt unter Umständen zu einer noch unklareren Abgrenzung zwischen den Tätigkeiten.

Des weiteren wird der zeitliche Aspekt der Expertise außer Acht gelassen. Eine Person, die im Extremfall vor vielen Jahren einige Zeit als Virologe tätig war, danach jedoch seinen Beruf gewechselt hat und zum Beispiel in die Politik gegangen ist, wird nicht mehr auf dem aktuellsten Stand der Forschung sein. Expertise kann sich im Laufe der Zeit verändern, was auch in [10] ausgeführt wird, indem die Aufgabe des *Temporal Expertise Profiling* eingeführt wird. Der in dieser Arbeit verwendete Expertise-Begriff unterscheidet nicht zwischen vergangener und derzeitiger Expertise (auf Basis der Tätigkeit), wodurch einer Person fälschlicherweise Expertise zugeschrieben werden könnte, obwohl sie schon lange nicht mehr in diesem Themengebiet tätig war.

Kapitel 4

Nächster-Zentroid-Klassifizierer

Um Tweets ihrem Inhalt nach einem Tätigkeitsfeld zuordnen zu können, soll ein Nächster-Zentroid-Klassifizierer entworfen werden, der mithilfe des Celebrity-Korpus trainiert und getestet wird. Anhand der Übereinstimmung des vorausgesagten Tätigkeitsfeldes eines untersuchten Tweets mit dem Tätigkeitsfeld, in dem der Autor entsprechend zuvor geschilderter Annahme Expertise aufweist, soll entschieden werden, ob der Tweet entfernt oder beibehalten wird. Ein Tweet zum maschinellen Lernen soll sichtbar bleiben, falls er durch einen Informatiker verfasst wird, jedoch nicht, falls der Autor ein Sportler oder Schauspieler ist. Abbildung 4.1 zeigt die generellen Schritte, die der entworfene Expertise-Filter ausführen muss, im Überblick.

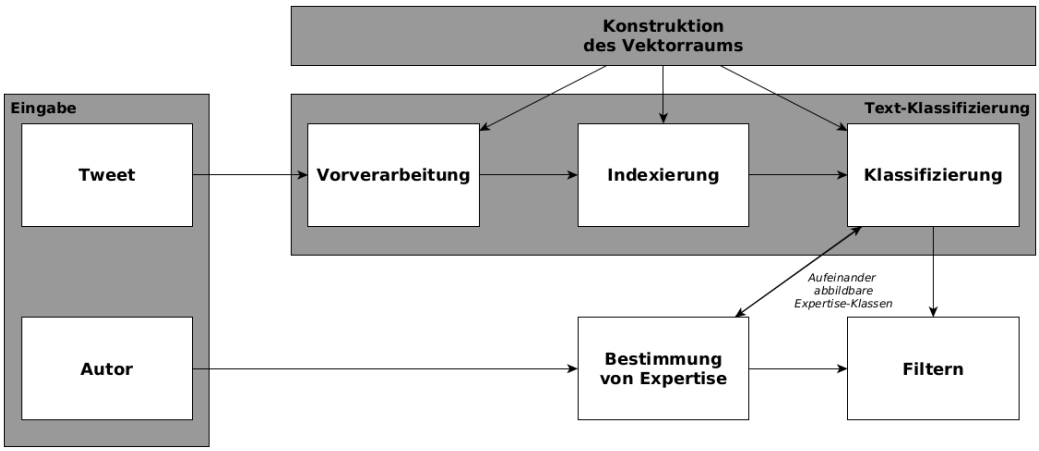


Abbildung 4.1: Pipeline des Expertise-Filters

Jede Tätigkeit soll durch einen Zentroiden in einem Vektorraum repräsentiert werden, der aus Vektorrepräsentationen der zugehörigen Tweets gebildet wird. Wie zuvor geschildert, wird jeder Tweet eines Autors, der nach dem

Celebrity-Korpus eine bestimmte Tätigkeit ausführt, ebendieser Tätigkeit zugeordnet. Hierfür muss ein solcher Vektorraum aufgebaut werden, in dem die Tweets eingebettet werden können. Dabei wird dem Bag-of-Words-Ansatz gefolgt, nach dem die Tweets als Multimenge ihrer enthaltenen Wörter repräsentiert werden und die Dimensionen des Vektorraums den Wörtern aus dem Vokabular des Korpus entsprechen. Nach einer Bereinigung und Normalisierung der Tweets werden die Vektoren mithilfe der Term-Häufigkeit und inversen Dokumenthäufigkeit (TFIDF) berechnet und anschließend L2-normalisiert. Das Ableiten von solchen messbaren Eigenschaften aus den Daten wird als Feature Extraction bezeichnet. Vektoren können miteinander verglichen werden, indem Distanzmaße genutzt werden. Ein Distanzmaß ist die Cosinusdistanz, das die Richtung, in die die beiden Vektoren zeigen, miteinander vergleicht. Dieses Maß wird angewendet, indem Tweets mit den gebildeten Tätigkeitszentroiden verglichen werden und der nächste Zentroid als die Klasse des Tweets ausgewählt wird.

Zur Auswertung des entwickelten Klassifizierers werden 200 zuvor separierte und manuell bewertete Test-Tweets als Eingabe für den Expertise-Filter verwendet. Die Ergebnisse zeigen, dass der Filter stark überfiltert, indem zu viele potentiell interessante Tweets entfernt werden. Insgesamt zeigt der Ansatz keine zufriedenstellenden Ergebnisse, wobei noch weitere Möglichkeiten zur Verbesserung diskutiert werden.

4.1 Repräsentation von Tweets im Vektorraum

In Bezug auf das Vektorraummodell werden Tweets als Dokumente D_i bezeichnet, die durch einen oder mehrere Terme t_j identifiziert werden. Dieser Ansatz der Textrepräsentation wird allgemein *Bag-of-Words* (BoW) genannt, da alle Worte eines Dokuments ungeordnet, also ungeachtet ihrer Wortreihenfolge, als Multimenge betrachtet werden. Diese Repräsentation ist einfach verständlich und schnell umzusetzen, hat jedoch zur Folge, dass der Kontext der Wörter ignoriert wird.

Die Dokumente aus der Dokumentensammlung, dem Celebrity-Korpus, werden in einem hochdimensionalen Vektorraum eingebettet. Vereinfacht kann man sich vorstellen, dass jedem Wort aus dem Vokabular des Korpus, das heißt alle n verschiedenen Wörter, die darin vorkommen, eine Dimension zugeordnet wird. Da Dokumente häufig unterschiedlicher Herkunft sind und in unterschiedlichen Formaten vorliegen oder sonstige strukturelle Unterschiede aufweisen, werden die Dokumente zuvor noch bereinigt und normalisiert, um einen einheitlichen Umgang mit ihnen zu gewährleisten.

4.1.1 Datenbereinigung und Normalisierung

In Kapitel 3.2 wurden einige Eigenschaften von Tweets beschrieben, die es erschweren, Data-Mining zu betreiben. Neben einer Vorauswahl geeigneter Tweets sind einige Schritte und Korrekturen notwendig, um Tweets zu normalisieren. In Abbildung 4.2 sind diese Schritte aufgezeigt.

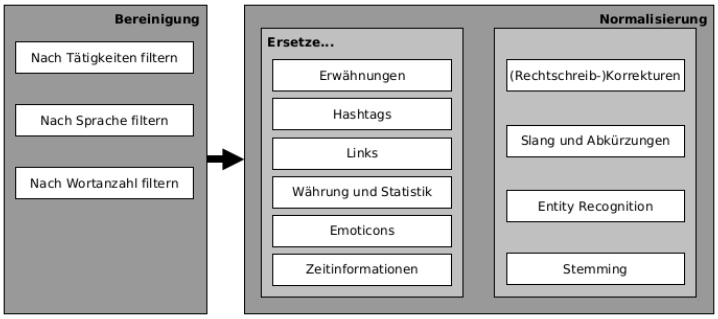


Abbildung 4.2: Schritte zur Bereinigung der Tweets

Wie zuvor beschrieben, muss der Korpus zunächst auf die Profile reduziert werden, die nur eine einzige Tätigkeit aufweisen. Die einem Prominenten zugeordneten Tätigkeiten sind als Liste hinterlegt, sodass nur diejenigen beibehalten werden, deren Listengröße genau 1 ist. Auch möglicherweise unzureichend ausgefüllte Profile ohne Tätigkeit werden so verworfen.

Als Stoppwörter einer Sprache werden Wörter bezeichnet, die sehr häufig in dieser Sprache vorkommen und keinen Informationsgehalt bieten. Beispiele aus dem Englischen sind „and“, „an“ oder „in“ und diese Worte können bedenkenlos aus den Tweets entfernt werden, wofür gängige Stoppwort-Listen herangezogen werden. Auch die Punktuation wird entfernt. Weiter werden Erwähnungen, Retweet-Symbole, Links, Zahlen und Emoticons durch die allgemeinen Bezeichner <user>, <retweet>, <link>, <hashtag>, <number> und <emoticon> ersetzt. Nach dieser Textersetzung werden die verbleibenden Worte gezählt (genannte Bezeichner ausgeschlossen) und Tweets mit nur einem oder sogar gar keinem verbliebenden Wort aus der Dokumentensammlung entfernt, da diese zu wenig Informationen enthalten und kategorisch als Nicht-Expertise-Tweets betrachtet werden können.

Mehrere verwendete Sprachen in den zugrundeliegenden Texten sind für die Textanalyse äußerst hinderlich und können nicht ohne weiteres in einen Vektorraum eingebettet werden. Um die betrachteten Tweets auf eine Sprache, Englisch, zu reduzieren, müssen alle fremdsprachigen Tweets aus dem Korpus entfernt werden. Dies geschieht mittels der in [6] beschriebenen Methode und Software *langid.py*, die mit hoher Genauigkeit die Sprache auch von kurzen

Texten identifizieren kann. Indem ein vortrainierter Naive-Bayes-Klassifizierer über Byte-n-gramme ($1 \leq n \leq 4$) angewendet wird, kann jedem Tweet seine Sprache zugeordnet werden. Dieser Klassifizierer wurde domänenunabhängig entwickelt, um ihn „von der Stange“ ohne notwendige Anpassungen auf den zugrundeliegenden Datensatz zu nutzen, und auch auf Texten von Plattformen wie Twitter getestet. Hierbei erreichte der Klassifizierer eine Treffergenauigkeit von 0.94. [6] Schließlich werden nur die Tweets beibehalten, die eindeutig als Englisch identifiziert wurden.

Zuletzt werden die Wörter der verbleibenden Tweets *gestemmt*. Dies bezeichnet die Reduktion verschiedener morphologischer Varianten eines Worts auf einen Wortstamm. Dies dient dazu, die inhaltlich gleichbedeutenden Varianten zu einem Merkmal zusammenzufassen. Das Stemming wird mithilfe des Porter-Stemmer-Algorithmus durchgeführt, der auf einer Menge von Verkürzungsregeln basiert. Die entstehenden Wortstämme sind nicht zwangsläufig linguistisch korrekt, jedoch wird das Ziel, die Zurückführung verwandter Wörter auf einen Stamm, dennoch erreicht. Das Ergebnis dieser Schritte ist eine Liste der bearbeiteten Wörter und Bezeichner, die allgemein als *Tokens* bezeichnet werden, woraus im nächsten Schritt Vektoren gebildet werden können.

Eine besondere Herausforderung im Umgang mit Kurztexten in sozialen Medien im Allgemeinen und Tweets im Speziellen ist - neben der geringen Wortanzahl - die Anwesenheit von Abkürzungen bzw. Slang, Rechtschreibfehlern und besonders Eigennamen.

4.1.2 Feature Extraction

Jeder Term aus dem nun reduzierten Vokabular bildet ein Merkmal der Vektoren, wobei nur eine Teilmenge der am häufigsten vorkommenden Terme beibehalten wird, um die entstehenden Modelle zu vereinfachen und Trainingszeiten zu reduzieren. Hier werden die 10000 Terme mit der höchsten Dokumenthäufigkeit beibehalten, wobei diese nicht in über 50 Prozent der Dokumente vorkommen dürfen. Terme, die in mehr als der Hälfte der Dokumente erscheinen, sind zur Unterscheidung dieser Dokumente höchstwahrscheinlich nicht brauchbar.

In Abbildung 4.3 ist zur einfachen Vorstellung ein dreidimensionaler Vektorraum aufgezeichnet. Dokumente nehmen in Abhängigkeit ihrer enthaltenen Terme unterschiedliche Positionen in diesem Raum ein. Terme können entweder entsprechend ihrer Relevanz gewichtet werden, oder ungewichtet in Abhängigkeit des Vorhandenseins des Terms im Dokument auf die binären Werte 0 und 1 beschränkt werden. Das am häufigsten verwendete Gewichtungsmodell TFIDF beruht auf der Term-Häufigkeit und der inversen Dokumenthäufigkeit

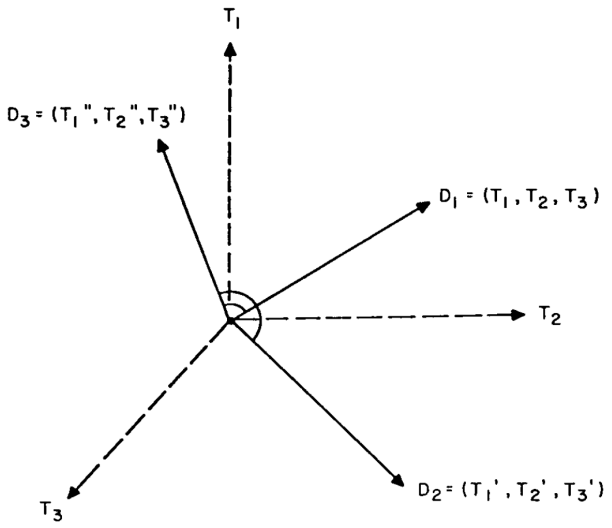


Abbildung 4.3: Vektorrepräsentation von Dokumenten wie in [11]

eines Terms t_j im Dokument D_i und wird wie folgt berechnet:

$$weight_{t_j,D_i} = \begin{cases} tf_{t_j,D_i} \cdot (log\frac{1+n}{1+x_{t_j}} + 1) & \text{if } tf_{t_j,D_i} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

wobei tf_{t_j,D_i} die Term-Häufigkeit eines Wortes t_j im Dokument D_i ist, n die Gesamtanzahl an Dokumenten im Korpus und x_{t_j} die Anzahl an Dokumenten, in denen das Wort t_j vorkommt. [13, 8] Die inverse Dokumenthäufigkeit misst die Spezifität eines Terms für alle betrachteten Dokumente und hängt somit nicht von einem einzelnen Dokument, sondern dem gesamten Korpus ab. Hier liegt der Gedanke zugrunde, dass dass Übereinstimmen von seltenen Begriffen aussagekräftiger ist als das Übereinstimmen von häufigen Begriffen.

Die entstandenen Vektoren werden im Anschluss L2-normalisiert, was auch als Euklidische Norm bezeichnet wird. Dabei wird der Vektor auf die Einheitslänge reduziert, indem er durch seine eigene Länge geteilt wird:

$$v_{L2} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + ... + v_n^2}}$$

Der konstruierte Vektorraum enthält nun die Vektorrepräsentationen aller relevanten Tweets.

4.2 Konstruktion des Klassifizierers

Alle Dokumente sind nun als Vektoren in den Vektorraum eingebettet. Insgesamt wurde eine Vorauswahl von 20 Tätigkeiten getroffen, sodass jedes Tweet-

Dokument einer dieser 20 Tätigkeiten zugeordnet ist. Diese Tätigkeiten sind die Zielklassen des Klassifizieres und die Dokumente das *Training Set*, auf dem der Klassifizierer trainiert wird.

Ein einfacher Klassifizierer repräsentiert nun jede Klasse, indem er den Zentroidvektor aller seiner enthaltenen Dokumente bildet, sodass wiederum 20 Zentroiden entstehen. Die Klasse eines neuen Dokuments x wird nun bestimmt, indem die Ähnlichkeiten zwischen x und allen 20 Zentroiden mithilfe der Cosinus-Ähnlichkeit berechnet wird und das Dokument dem ähnlichsten Zentroiden und der damit verbundenen Klasse zugeordnet wird. [16] Im Vektorraummodell wird die Ähnlichkeit zwischen zwei Dokumenten D_i und D_j gewöhnlich mithilfe der Cosinus-Funktion berechnet, die definiert ist als

$$cos(D_i, D_j) = \frac{D_i \cdot D_j}{\|D_i\|_2 * \|D_j\|_2}$$

wobei $\cdot$ das Skalarprodukt zweier Vektoren bezeichnet. Durch die Normalisierung der Vektoren auf Einheitslänge vereinfacht sich die Berechnung auf $cos(D_i, D_j) = D_i \cdot D_j$.

Ausgehend von einer Menge S_T an Dokumenten, die einer Tätigkeit T zugeordnet sind, wird der Zentroidvektor C_T dieser Tätigkeit definiert als

$$C_T = \frac{1}{|S_T|} \sum_{D \in S_T} D$$

was dem Mittel der verschiedenen Termgewichtungen in den Dokumenten von S_T entspricht. Die Ähnlichkeit zwischen einem Dokument und einem Zentroidvektor kann schließlich wie folgt berechnet werden:

$$cos(D, C_T) = \frac{D \cdot C_T}{\|D\|_2 * \|C_T\|_2} = \frac{D \cdot C_T}{\|C_T\|_2}$$

Nach der Berechnung aller k Zentroiden $\{C_1, C_2, ..., C_k\}$ in der Trainingsphase, wobei C_i der Zentroid der i-ten Tätigkeitsklasse ist, kann die Klasse eines neuen Dokuments x bestimmt werden, indem die Ähnlichkeit zwischen zu allen Zentroiden bestimmt wird und der Ähnlichste als Klasse gewählt wird. Dies entspricht folgender Optimierung:

$$arg \max_{j=1,...,k} (cos(x, C_j))$$

Im Ergebnis wird durch diesen Klassifizierer ein Dokument eine der 20 Expertise-Klassen zugeordnet, die zuvor aus dem Celebrity-Korpus extrahiert wurden. Generell müssen die Ausgaben des Klassifizierers auf die Expertise-Klassen der Autoren abgebildet werden können, die durch Methoden des Expertise Profiling bestimmt werden.

4.3 Auswertung

Der entworfene Nächster-Zentroid-Klassifizierer wurde auf einem *Test Set* von insgesamt 200 zuvor separierten und manuell ausgewerteten Tweets angewendet, um die Performanz des Klassifizierers auszuwerten. Hier enthalten sind 113 Expertise-Tweets und 87 Nicht-Expertise-Tweets. Zur Evaluation werden die Maße „Treffergenauigkeit“ (engl. Accuracy), „Sensitivität“ (engl. Recall) und der „Negative Vorhersagewert“ (engl. negative predictive value) betrachtet.

Die Treffergenauigkeit gibt allgemein an, in wie vielen Fällen der Klassifizierer korrekt klassifiziert. Die Sensitivität hingegen gibt den Anteil der korrekt als positiv klassifizierten Objekte an der Gesamtheit der tatsächlich positiven Objekte an, hier also den Prozentsatz der tatsächlichen Expertise-Tweets, die durch den Klassifizierer erhalten wurden. Dies ist eine besonders wichtige Eigenschaft, da der Verlust an wertvollen Informationen durch den Filter minimiert werden soll. Der negative Vorhersagewert gibt den Anteil der korrekt als negativ klassifizierten Ergebnisse an der Gesamtheit der als negativ klassifizierten Ergebnisse an und gibt somit Auskunft über das Verhältnis, wieviele Nicht-Expertise-Tweets zu Ungunsten von Expertise-Tweets entfernt wurden. Wünschenswert sind hier Ergebnisse über 0.5, da somit mehr Nicht-Expertise-Tweets als Expertise-Tweets entfernt würden, sodass in der Timeline des Nutzers verhältnismäßig mehr Expertise-Tweets sichtbar wären.

$$\text{Treffergenauigkeit} = \frac{r_p + r_n}{r_p + r_n + f_p + f_n}$$

$$\text{Sensitivität} = \frac{r_p}{r_p + f_n}$$

$$\text{Negativer Vorhersagewert} = \frac{r_n}{r_n + f_n}$$

Um die Qualität des entworfenen Klassifizierers einzuordnen, wird dieser mit einer Baseline verglichen. Als Baseline kann der triviale Fall angenommen werden, dass der Expertise-Filter keinen Tweet aus der Timeline entfernen würde, also alle Tweets als Expertise-Tweets einordnet. Somit hätte diese Baseline eine Treffergenauigkeit von 0.565 und eine Sensitivität von 1.0. Der negative Vorhersagewert kann nicht berechnet werden, da keine Tweets als Nicht-Expertise-Tweets klassifiziert werden, und wird daher als 0 definiert.

Wie in Abbildung 4.4a und 4.4b ersichtlich, weist der Nächster-Zentroid-Klassifizierer insgesamt eine schlechtere Treffergenauigkeit als die Baseline auf. Zudem gehen über 70% der Expertise-Tweets verloren, was eine äußerst starke Überfilterung anzeigt. Der negative Vorhersagewert bleibt unter 0.5, was zeigt, dass insgesamt auch mehr Expertise-Tweets als Nicht-Expertise-Tweets aus der Timeline entfernt werden würden. Die Konfusionsmatrix zeigt die Ergebnisse

Wahr \ Filter	E	N
E	33	80
N	12	75

(a) Konfusionsmatrix

Maß	Baseline	NCC
Treffergenauigkeit	0.565	0.54
Sensitivität	1.0	0.292
NPV	0	0.484

(b) Statistische Auswertung

Abbildung 4.4: Auswertung: Nächster-Zentroid-Klassifizierer

in absoluten Zahlen, wobei E für „Expertise“ und N für „Nicht-Expertise“ steht.

Die Ergebnisse sind insbesondere darauf zurückzuführen, dass die gebildeten Zentroiden eine große Anzahl an Nicht-Expertise-Tweets enthalten. Wie zuvor geschildert, umfassen die veröffentlichen Tweets zu ungefähr einer Hälfte Nicht-Expertise-Tweets, die aufgrund derzeitig mangelnder Unterscheidungsmöglichkeiten auch in die Bildung der Zentroiden der unterschiedlichen Tätigkeiten einbezogen werden. Dies sorgt dafür, dass die Zentroiden einander ähnlich sind und nur schwierig voneinander zu trennen.

Zudem existiert keine „Nicht-Expertise“-Klasse, der solche Tweets zugeordnet werden könnten, die generell keinen Anspruch auf Expertise haben und zum Beispiel in die Kategorie privater Nachrichten fallen. Derzeit werden sie zufällig einem anderen Zentroiden zugeordnet, die alle solche Nicht-Expertise-Tweets enthalten. So kann beispielsweise auch eine Nachricht eines Informatiker, die normalerweise als Nicht-Expertise erkannt werden sollte, der Informatiker-Tätigkeit zugeordnet werden, sodass sie fälscherlicherweise als Expertise-Tweet beibehalten wird. Eine Lösung dieses Problems könnte sein, eine Vorklassifizierung durchzuführen, die Nicht-Expertise-Tweets zu einer Klasse zusammenfasst und daraus einen Zentroiden bildet.

Insbesondere die Vorverarbeitung der Tweets ist derzeit nur unzureichend behandelt. Wichtige Korrekturen wie das Entfernen von Slang bzw. Abkürzungen, Rechtschreibfehlern und Wortbetonungen sind bisher nicht umgesetzt, sodass das Vokabular noch einige Redundanzen enthält, die nicht auf einen Term zurückgeführt werden. Auch die Erkennung von Entitäten in den Tweets ist nicht gewährleistet, weshalb wertvolle Eigennamen nicht als solche in die Analyse einfließen. Besonders auf Twitter, das aktuelle Neuigkeiten auf Basis von Personen, Organisationen, Orten und dergleichen bespricht, ist dies ein schwerwiegendes Problem.

Kapitel 5

Weitere Ansätze zur Klassifizierung

Aus der Auswertung des Nächster-Zentroid-Klassifizierers sind zwei große Probleme ersichtlich geworden, die die schlechte Performanz des Klassifizierers zumindest teilweise erklären:

- Die Zentroiden sind durch zu viele Nicht-Expertise-Tweets verunreinigt und daher nur sehr schwierig voneinander zu trennen.
- Nicht-Expertise-Tweets können nicht verlässlich klassifiziert werden, da keine solche Klasse existiert und auch nicht ohne Weiteres im Voraus aus den Daten abgeleitet werden kann.

In Folge dessen werden zwei weitere Ansätze verfolgt, die jeweils eines dieser Probleme behandeln. Zum einen wurde der vorhandene Klassifizierer so angepasst, dass die Zentroiden nun nicht mehr aus den Tweets von Autoren der entsprechenden Tätigkeiten gebildet wurden, sondern aus Wikipedia-Artikeln, die diese Tätigkeiten näher beschreiben. Dies umgeht die Problematik der großen Anzahl an Nicht-Expertise-Tweets, da die Wikipedia-Artikel in jedem Fall nur für die Tätigkeit relevante Informationen enthalten, sodass hier ein ausgeprägteres und spezielleres Vokabular entsteht, das die Tätigkeiten schärfer voneinander abgrenzen soll. Zum anderen wurde eine *OneClass Support Vector Machine* (*SVM*) auf einer allgemeinen Expertise-Klasse konstruiert, für die 1000 Tweets per Hand ausgewertet wurden, wovon 483 erkannte Expertise-Tweets als Trainingsdaten dienen. Hiermit sollen generell Expertise-Tweets von Nicht-Expertise-Tweets voneinander getrennt werden können, ohne Rücksicht auf die Expertise des Autoren zu nehmen.

Sowohl der Wikipedia-basierte Klassifizierer als auch die konstruierte SVM weisen eine sehr geringe Treffergenauigkeit auf. Somit können beide Ansätze

die eingangs genannte Problematik nicht lösen. Während die OneClass-SVM insgesamt konservativer verfährt und vergleichsweise weniger Tweets entfernt, zeigt der Wikipedia-basierter Klassifizierer eine noch stärkere Überfilterung als der ursprüngliche Tweet-basierte Klassifizierer.

5.1 Zentroiden aus Wikipedia-Artikeln

Die im Celebrity-Korpus angegebenen Tätigkeiten der einzelnen Autoren basieren auf Informationen von Wikidata. Zu jeder Tätigkeit, die in Wikidata als eigenes Objekt repräsentiert ist, existieren weitere Eigenschaften und Verknüpfungen. So lassen sich Wikipedia-Artikel zu den Tätigkeiten einfach finden oder - falls diese nicht vorhanden oder nur unzureichend sind - Verknüpfungen zu Spezialisierungen oder Generalisierungen der Tätigkeit. Außerdem weisen viele Tätigkeiten auch ein Objekt „Tätigkeitsfeld“ (engl. „field of occupation“) auf, das nicht nur den Beruf, sondern auch das relevante Themengebiet umreißt.

Die verlinkten Artikel umfassen inhaltlich eine Beschreibung der Tätigkeit, ihre Geschichte und Rolle im Laufe der Zeit, weitere (Wissens-)Felder in dem Tätigkeitsgebiet und speziellere, themenspezifische Kategorien von Informationen. So wird im Feld der Musik auch eine Vielzahl an Instrumenten vorgestellt oder in der Informatik unterschiedliche Programmierparadigmen. Zu jeder Tätigkeit fallen relevante Schlagworte, die (nur) für dieses Gebiet relevant sind, sodass aus der Dokumentensammlung ein umfangreiches Vokabular extrahiert werden kann.

Zu den 20 Tätigkeiten wurden, soweit vorhanden, also jeweils Wikipedia-Artikel sowohl zur Tätigkeit selbst als auch zum Tätigkeitsfeld in einzelnen Dokumenten zusammengefasst. Die entstandenen Dokumente wurden genauso vorverarbeitet wie die Tweets, genauer:

1. Elemente wie Emoticons, Links, Zahlen durch Repräsentaten ersetzen
2. Tokenisierung anhand von Leerzeichen
3. Entfernung von Stoppwörtern
4. Stemming mithilfe des Porter-Stemmer-Algorithmus

Aufgrund der Natur und des Verwendungszwecks von Wikipedia weisen die dort verfassten Texte andere stilistische Merkmale auf als die manchmal spontan veröffentlichten Tweets auf Twitter. Im Gegensatz zu Tweets weisen die Wikipedia-Artikel keinen Slang oder Rechtschreibfehler auf, werden nicht mit Wortabwandlungen betont oder durch umgangssprachliche Abkürzungen verändert. Auch Emoticons finden keine Verwendung, da die Artikel versuchen,

Wahr \ Filter	E	N
E	22	91
N	3	84

(a) Konfusionsmatrix

Maß	Baseline	NCC	NCC-Wiki
Treffergenauigkeit	0.565	0.54	0.53
Sensitivität	1.0	0.292	0.195
NPV	0	0.484	0.48

(b) Statistische Auswertung

Abbildung 5.1: Auswertung: Nächster-Zentroid-Klassifizierer (Wikipedia)

nur sachliche Beschreibungen zu liefern und keine Meinungen darzustellen. Diese stilistische Hürde führt dazu, dass Informationen der Tweets beim Einbetten in einen Vektorraum, der mithilfe der Wikipedia-Artikel konstruiert wurde, verloren gehen können. Diese Problematik muss mit weitergehenden Vorverarbeitungsschritten, die genannte Abwandlungen in den Tweets wieder normalisieren, begegnet werden, was jedoch in dieser Arbeit nicht mehr behandelt wird. Auf der anderen Seite liegt nahe, dass zumindest diejenigen Tweets, die solche Anomalien wie Slang bzw. Abkürzungen besitzen, entweder gar keine Expertise-Tweets sind oder das entscheidene Vokabular unberührt lassen, da solche mit mehr Zeit und Bedacht verfasst werden.

Die Ergebnisse zeigen, dass der auf Wikipedia-Artikeln basierende Nächster-Zentroid-Klassifizierer bei allen angewendeten Maßen schlechter abschneidet als der ursprüngliche Klassifizierer und damit auch als die Baseline. Die falschpositive Rate ist nun reduziert, was sich möglicherweise darauf zurückführen lässt, dass mit der Spezialisierung des Vokabulars die Nicht-Expertise-Tweets nicht mehr „zufällig“ einer Klasse zugeordnet werden. Die stärkere Überfilterung hat seine Ursache ebenfalls in der Änderung und Spezialisierung des Vokabulars.

Die Vermutung liegt nahe, dass die Aussagen auf Twitter, die tatsächlich von Interesse sind, weniger von fachspezifischem Vokabular abhängig sind, sondern mehr von aktuell relevanten Eigennamen, die in Wikipedia-Artikeln, die vor allem Historie und allgemeine Themenbeschreibungen abbilden, gar nicht enthalten sind. Kommentare zu wichtigen Konferenzen, zu speziellen Personen, Unternehmen oder Produkten werden nicht erfasst, da weder Entitäten erkannt werden noch diese den Zentroiden zugeordnet wurden. So kann der Inhalt vieler Expertise-Tweets nicht korrekt eingeordnet werden, weshalb diese dann als Nicht-Expertise verworfen werden.

Diese Problematik bleibt selbst bei einem erfolgreichen Erkennen der Entitäten erhalten, da der Klassifizierer nur auf einen Ausschnitt der relevanten Entitäten trainiert wird und neue Entitäten, die im Laufe der Zeit hinzukommen oder relevant werden, nicht erkannt werden. Die schwach besseren Ergebnisse des ursprünglichen Nächster-Zentroid-Klassifizierers lassen sich auch damit

erklären, dass sowohl Trainings- als auch Testdaten aus demselben Zeitfenster stammen und somit grundsätzlich gleiche oder ähnliche Themen ansprechen.

5.2 SVM zur Bestimmung von Expertise-Tweets

Aus der Auswertung des Nächster-Zentroid-Klassifizierers ergibt sich die Einschätzung, dass eine Nicht-Expertise-Klasse notwendig sei, um Tweets nicht nur nach ihrer Tätigkeit zuordnen zu können, sondern auch solche Tweets zu erfassen, die sich thematisch gar nicht mit einem Tätigkeitsfeld auseinandersetzen. Es soll ein binärer Klassifizierer entworfen werden, der grundsätzlich unterscheiden kann zwischen Expertise-Tweets und Nicht-Expertise-Tweets, jedoch nicht, ob das Themenfeld auch mit der Expertise des Autoren übereinstimmt. Dieser Klassifizierer kann verwendet werden, um im Voraus zu entscheiden, welche Tweets weitergehend untersucht werden sollen.

Zur Realisierung eines solchen Klassifizierers wird eine OneClass Support Vector Machine nach Schölkopf et. al. [7] auf 483 manuell bestimmten Expertise-Tweets (von insgesamt 1000 ausgewerteten Tweets) trainiert und 200 separat ausgewerteten Tweets getestet. Dabei liegt die Idee zugrunde, dass sich Expertise-Tweets Merkmale teilen, die von der SVM erfasst werden können. Diese Merkmale sind hier Worte des Vokabulars und allgemein ist das Vokabular von Expertise-Tweets spezieller und ausgeprägter, da sie sich tiefergehend mit einem Thema auseinandersetzen.

Die OneClass Support Vector Machine versucht zu entscheiden, ob ein neuer Datenpunkt zu einer Klasse, die durch die Trainingsdaten definiert wird, gehört oder zu „verschieden“ von dieser Klasse ist und ihr somit nicht zuzuordnen ist. Die OneClass-SVM wird dann verwendet, wenn sich Daten zu einer Klasse einfach akquirieren lassen, jedoch Gegenbeispiele einer anderen Klasse nur schwierig oder nicht umfassend genug finden lassen. Es werden also nur positive Beispiele einer Klasse zum Training verwendet.

Im Wesentlichen wird eine Funktion f berechnet, die einen Merkmalsraum so teilt, dass Elemente der Expertise-Klasse E positiv und die ihres Komplements, der Nicht-Expertise-Klasse N, negativ abgebildet werden.

$$f(x) = \begin{cases} +1 & \text{if } x \in E \\ -1 & \text{if } x \in N \end{cases}$$

Die kleine Region der Klasse E soll dabei möglichst alle Vektoren der Expertise-Tweets umfassen. Die SVM separiert alle Datenpunkte vom Ursprung und maximiert dabei die Distanz von der entstandenen Hyperebene zum Ursprung. Da die Berechnung der Distanz nur vom Skalarprodukt der Vektoren abhängig ist, kann eine Kernel-Funktion gewählt werden, die dieselben Ergebnisse für

		Filter	
		E	*N*
Wahr	*E*	72	41
	N	60	27

(a) Konfusionsmatrix

Maß	Baseline	SVM
Treffergenauigkeit	0.565	0.495
Sensitivität	1.0	0.637
NPV	0	0.484

(b) Statistische Auswertung

Abbildung 5.2: Auswertung: Support Vector Machine

diese Berechnung liefert, aber keine Projektion zum hochdimensionalen Merkmalsraum erfordert. Die Funktion f entscheidet dann, ob ein Vektor zur Klasse E gehört oder nicht.

Die Ergebnisse der konstruierten OneClass-SVM sind nicht vielversprechend. Während die SVM zwar bedeutend mehr Tweets erhält als die anderen Klassifizierer, weist sie insgesamt die geringste Treffergenauigkeit und ebenfalls einen negativen Vorhersagewert unter 0.5 auf. Auch die geringe Erkennungsrate der Nicht-Expertise-Tweets (nur ungefähr ein Viertel wurde erkannt) zeigt, dass der Klassifizierer in dieser Form nicht geeignet ist, eine Vorklassifizierung von Expertise-Tweets und Nicht-Expertise-Tweets zu leisten.

Es kann auf Basis der Trainingsdaten keine Entscheidungsfunkion gefunden werden, die die Expertise-Klasse zuverlässig beschreibt. Auch wenn die Support Vector Machine im Allgemeinen auch auf kleinen Datensätzen gute Ergebnisse zeigt, muss hier jedoch betont werden, dass die ausgewählten 483 Expertise-Tweets nicht stellvertretend für die gesamte Klasse stehen können. Dies wird insbesondere dadurch verschärft, dass hier dem Bag-of-Words-Ansatz gefolgt wird und nur ein sehr geringer Teil des relevanten Vokabulars erfasst wird. Da das Vokabular über die Expertise-Themenfelder hinweg sehr varianten- und umfangreich ist, muss überlegt werden, ob dieser Ansatz hier zielführend ist. Da gerade nicht das Thema der Tweets, sondern ihre stilistischen Eigenschaften und Strukur erfasst werden soll, wäre ein Ansatz auf Basis der in Kapitel 3 vorgestellten Tweet-Eigenschaften vorstellbar.

Kapitel 6

Fazit und zukünftige Arbeit

Im Rahmen dieser Arbeit konnte der Begriff der Expertise näher definiert werden, wobei begründet wurde, warum die Tätigkeit einer Person als Hinweis auf das Vorhandensein einer solchen Expertise zu bestimmten Themenfeldern dienen kann. Dabei wurden Probleme des Expertise-Profilings auf Basis von Tweets angesprochen. Expertise lässt sich nur schwerlich aus den reinen Tweets auf Twitter bestimmen, sondern erfordert im besten Fall zusätzlich externe Quellen wie die Zugehörigkeit zu thematischen Listen. Hieraus wurde die praktische Notwendigkeit einer umfassenden Datenbasis gefolgert, die sowohl umfangreiche Twitter-Profile als auch Informationen zu ihrer Expertise enthält. Ein geeigneter Korpus ist der Celebrity-Korpus, auf Basis dessen verschiedene Klassifizierer trainiert und getestet werden können.

Die konstruierten Klassifizierer zeigen keine aussichtsreichen Ergebnisse. Die Nächster-Zentroid-Klassifizierer, sowohl auf Basis von Tweets als auch von Wikipedia-Artikeln, leisten zwar in einigen Fällen durchaus richtige Zuweisungen, haben jedoch eine geringere Treffergenauigkeit als die Baseline und weisen zudem eine starke Überfilterung auf, die besonders für die Aufgabe eines konservativen Filters kritisch ist. Insgesamt werden bedeutend mehr potenziell interessante Expertise-Tweets entfernt als weniger interessante Nicht-Expertise-Tweets.

Die mangelnde Genauigkeit der Klassifizierer kann mit mehreren Faktoren erklärt werden. Zum einen beinhalten die erstellten Zentroiden eine zu große Menge an normalerweise zu entfernenden Nicht-Expertise-Tweets, die die Tätigkeit nicht beschreiben können und somit die Repräsentation verzerren. Zum anderen existiert keine Nicht-Expertise-Klasse, der entsprechende Tweets zugeordnet werden können, anstatt sie „zufällig“ anderen Zentroiden zuzuordnen. Zukünftig soll versucht werden, eine umfangreichere Vorverarbeitung der Tweets voranzustellen, um so eine Nicht-Expertise-Klasse zu definieren und nicht relevante Kategorien von Tweets auszuschließen.

Der Versuch, eine solche Vorklassifizierung mithilfe der Support Vector Machine zu leisten, ist gescheitert. Während die OneClass-SVM zur allgemeinen Bestimmung von Expertise-Tweets eine merkbar höhere Sensitivität aufweist, ist seine Treffergenauigkeit die schlechteste und verbleibt ebenfalls unter der Baseline. Insgesamt ist auch hier kein Mehrwert sichtbar, da ebenfalls mehr Expertise-Tweets entfernt werden als Nicht-Expertise-Tweets und auch nur ein Bruchteil der Nicht-Expertise-Tweets als solche erkannt wurden. Problem der SVM ist, dass diese auf zuvor bestimmten Expertise-Tweets trainiert werden muss. Da dies manuell geschieht und somit nur eine sehr begrenzte Auswahl solcher Tweets zur Verfügung steht, kann die Expertise-Klasse nicht ausreichend erfasst werden.

Im Zuge der manuellen Untersuchung der Tweets konnten einige Kategorien von Tweets ermittelt werden, die auch mit der Literatur [14] übereinstimmen. Diese umfassen Neuigkeiten, Meinungen, Deals, Events und private Nachrichten. Dabei wurden einige Eigenschaften genannt, die zur Unterscheidung von Tweets herangezogen werden können, ohne deren thematischen Inhalt zu kennen. Das Vorhandensein bestimmter Elemente wie Links, Erwähnungen, Hashtags und Emoticons oder das Auftauchen von vokabularfremden Wörtern kann in zukünftiger Arbeit für Experimente genutzt werden, um Tweets im Voraus zu kategorisieren. Ebendiese Eigenschaften können auch als weiterer Ansatz verwendet werden, um die SVM zu verbessern, indem nicht mehr das (zeitabhängige) Vokabular als Merkmale der Repräsentation gebraucht werden, sondern diese generellen Eigenschaften.

Ein weiteres Problemfeld ist die Normalisierung von Tweets. Sie weisen viele Unregelmäßigkeiten wie Rechtschreibfehler, Slang bzw. Abkürzungen und Wortbetonungen auf, die zu vokabularfremden Wörtern führen. Dies betont die Notwendigkeit einer umfassenden Vorverarbeitung bei zukünftiger Arbeit, um diese Wörter wieder auf grammatikalisch korrekte Formen zurückzuführen. Auch Eigennamen bzw. Entitäten spielen eine große Rolle in Bezug auf das Vokabular. Ihre Erkennung und der Umgang damit muss bei zukünftiger Arbeit näher untersucht werden, da Neuigkeiten und Meinungen auf Twitter häufig Bezug auf solche Entitäten nehmen.

Insgesamt weisen die Experimente und die theoretische Betrachtung darauf hin, dass eine umfassende Vorverarbeitung und Säuberung der Daten notwendig ist, um Kurztexte wie Tweets thematisch zu klassifizieren. Auch wenn die der Bag-of-Words-Ansatz hier möglicherweise limitiert ist, sind die Grenzen der Ansätze noch nicht ausgeschöpft. Zu beachten ist jedoch besonders auch, dass sich die in den Expertise-Feldern angesprochenen Themen ändern und hierfür ein entsprechender Umgang gefunden werden muss. Jenseits des Bag-of-Words-Ansatzes sollten besonders die vorgestellten Eigenschaften von Tweets genauer untersucht und auf ihre Verwendbarkeit geprüft werden.

Literaturverzeichnis

[1] Krisztian Balog, Toine Bogers, Leif Azzopardi, Maarten de Rijke and Antal van den Bosch. Broad Expertise Retrieval in Sparse Data Environments. In *SIGIR 2007 Proceedings*, 2007.

[2] Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov and Luo Si. Expertise Retrieval. In *Foundations and Trends® in Information Retrieval: Vol. 6: No. 2–3*, July 2012, pp 127-256.

[3] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam und Ed H. Chi. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *UIST'10*, Oktober, 2010, New York, USA.

[4] Eui-Hong (Sam) Han and George Karypis. Centroid-Based Document Classification: Analysis and Experimental Results. In *PKDD 2000, LNAI 1910*, pp. 424–431, 2000.

[5] Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, June, 2011, pp. 368–378.

[6] Marco Lui and Timothy Baldwin. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Republic of Korea, July, 2012, pp. 25-30

[7] Larry M. Manevitz and Malik Yousef. One-Class SVMs for Document Classification. In *Journal of Machine Learning Research 2*, 2001, pp. 139-154.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research, Vol. 12*, 2011, pp. 2825-2830.

[9] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman and Robert Frederking. Topical Clustering of Tweets. In *SWSM'10*, Beijing, China, Juli 2011.

[10] Jan Rybak, Krisztian Balog und Kjetil Nørvåg. Temporal Expertise Profiling. In *European Conference on Information Retrieval*, Springer, Cham, 2014. S. 540-546.

[11] G. Salton, A. Wong and C. S. Yang. A Vector Space Model for Automatic Indexing. In *Communications of the ACM, Vol. 18: No. 11*, November 1975, pp. 613-620.

[12] Naveen Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly und Krishna P. Gummadi. Inferring Who-is-Who in the Twitter Social Network. In *WOSN'12*, Helsinki, Finland, August, 2012.

[13] P. Soucy, G. W. Mineau. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. In *IJCAI, Vol. 5*, Juli 2005, pp. 1130-1135.

[14] Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas. Short Text Classification in Twitter to Improve Information Filtering. In *SIGIR'10*, Geneva, Switzerland, Juli 2010.

[15] Matti Wiegmann, Benno Stein, Martin Potthast. Celebrity Profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, July 2019.

[16] Vandana Korde and C Namrata Mahender. Text Classification and Classifiers: A Survey. In *IJAIA, Vol.3, No.2,* März 2012.

[17] C. Wagner, V. Liao, P. Pirolli, L. Nelson and M. Strohmaier. It's Not in Their Tweets: Modeling Topical Expertise of Twitter Users. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Amsterdam, September 2012, pp. 91-100.