Text Mining für Nachrichten

Joachim Schole
Hochschule für angewandte
Wissenschaften Hamburg
Fakultät Technik und Informatik
Department Informatik
Email: joachim.schole@haw-hamburg.de

I. Einführung

Diese Arbeit befasst sich mit dem Thema des Text Mining im Nachrichtenbereich. Text Mining gilt weithin als Unterkategorie des Data Mining [1]. Der Begriff Data Mining beschreibt den Prozess, Informationen aus strukturierten Daten zu ermitteln. Das Ziel ist dabei, neue Informationen zu erlangen, welche bisher nicht zur Verfügung standen [2]. Dies können sogar Informationen sein, nach denen nicht explizit gesucht wurde. Text Mining beschreibt eine spezielle Ausprägung des Data Mining, welches sich auf die Auswertung von Textdaten beschränkt [1]. Es existieren viele unterschiedliche Forschungsansätze und -Arbeiten im Bereich des Text Mining. Im Bereich der Nachrichten liegt der Fokus unter anderem auf der automatischen Erkennung und Zusammenfassung von Ereignissen. Außerdem wird versucht, Nachrichtenarchive mittels Text-Mining-Methoden aufzuarbeiten, um beispielsweise Dossiers zu erstellen.

Zunächst führt diese Arbeit eine Begriffsdefinition zum Text Mining durch. Daraufhin werden Ansätze zur Ereignis- und Themenzusammenfassung unter Verwendung von Twitter-Daten vorgestellt. Weiter führt diese Arbeit in die Thematik der Dossiererstellung ein. Abschließend gibt die Arbeit einen Ausblick auf mögliche Arbeiten für das kommende Grundprojekt.

II. Schnittstellen zu anderen Projekten

Es existieren einige Schnittstellen zu anderen Arbeiten an der HAW Hamburg. Am höchsten ist die Überschneidung zu den Arbeiten von Hälker [3] [4]. Diese befassen sich mit den Themen des Information Retrieval [4] und des Text Minings im Nachrichtenbereich [3]. Weiter sind die Arbeiten von Schöneberg [5] [6] [7] relevant. Diese befassen sich mit der Ermittlung von *Weak Signals* im Hinblick auf die Früherkennung von Trends [5], der Trenderkennung in Texten [6] und der Erstellung von Dossiers [7]. With [8] vergleicht in seiner Arbeit verschiedene Text-Mining-Algorithmen im Hinblick auf die Eignung für die Verwendung in einem Epidemie-Frühwarnsystem. Da diese Arbeit sich mit Nachrichten befasst, ist der gemeinsame Bezug an dieser Stelle eher gering. Demin [9] [10] verwendet ebenfalls Text Mining. Allerdings liegt hier der Fokus auf der Entwicklung einer Second-Screen-Anwendung.

III. Hauptteil

Dieses Kapitel stellt einige Forschungsarbeiten zu Text-Mining-Themen vor. Zunächst findet eine Begriffserklärung des Text Mining und eine Abgrenzung zum Data Mining dar. Als Grundlage hierzu dienen die Arbeiten von Hipper und Rentzmann [1] und Kroeze et al. [2]. Anzumerken ist hier, dass letztere bereits Hälker [3] eingehend behandelt. Daraufhin erfolgt eine Darstellung einiger Forschungsarbeiten zum Thema Nachrichten in sozialen Medien mit Fokus auf Twitter als Datenquelle. Als Grundlage hierzu dienen die Arbeiten von Chua und Asur [11], Zhang et al. [12] und Lehmann et al. [13] [14]. Weiter erfolgt eine Einführung in das Thema der Dossiererstellung.

A. Begriffserklärung Text Mining

Der Prozess des Text Mining ähnelt dem des Data Mining. Der größte Unterschied liegt in der verwendeten Datenquelle. Während beim Data Mining strukturierte Daten beispielsweise in Form einer relationalen Datenbank die Grundlage bilden, finden beim Text Mining Textdokumente Verwendung [2]. Diese werden allgemein als unstrukturierte, beziehungsweise semi-strukturierte [15] Daten bezeichnet. Dies ergibt sich daraus, dass in Textdokumenten in Form der Grammatik eine implizite Struktur vorhanden ist. Weiter verfügen einige Textdokumente in Form von Kapiteln und anderer Unterteilung über eine explizite Struktur [1].

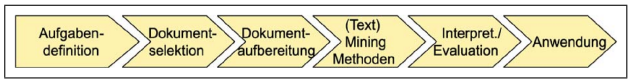


Figure 1. Text Mining Prozess nach Hippner und Rentzmann [1, S. 288]

Um Data Mining Methoden auf Textdokumente anwenden zu können, müssen die zugrundeliegenden Textdokumente zunächst aufbereitet werden. Hierfür bestehen verschiedene Möglichkeiten. Je nach angestrebter Weiterverarbeitung muss hier eine geeignete Vorgehensweise gefunden werden. So bestehen beispielsweise die Möglichkeiten, eine *morphologische*, *syntaktische* oder *semantische Analyse* durchzuführen [1]. Die morphologische Analyse untersucht ein Dokument im Hinblick auf die vorhandenen Wortformen. Dabei führt ein Algorithmus die vorhandenen Wörter auf ihren Stamm zurück. Dieses Verfahren heißt *Stemming*. Alternativ ist die

Vorgehensweise bei der *Lemmatisation*, die unterschiedlichen Formen eines Wortes unter Zuhilfenahme einer *Lookup-Table* auf eine einheitliche Form zu bringen. Manning et al. [16] definieren die Lemmatisation wie folgt:

"Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma." [16, S. 32]

Die syntaktische Analyse untersucht ein Dokument im Hinblick auf seinen grammatikalischen Aufbau. Das bekannteste Vorgehen hierfür ist das *Part-of-Speech Tagging*. Ein solcher Tagger versieht alle Wörter eines Dokuments mit Tags, welche die jeweilige Wortart angeben [1]. Dies geschieht ebenfalls unter Verwendung einer Lookup-Table. Die semantische Analyse betrachtet die Wörter eines Dokuments im Hinblick auf ihre kontextuelle Bedeutung. So kann besonders für mehrdeutige Wörter eindeutig eine Bedeutung ermittelt werden. Durch Anwendung dieser Methoden erhält ein Textdokument eine Struktur, welche weitere Algorithmen verarbeiten können. Beispiele für diese Algorithmen sind das häufig verwendete Identifizieren und Entfernen von *Stopwords* [17] und das anschließende Zählen von häufig vorkommenden Wörtern und Termen. Die so ermittelten häufigsten Terme können Aufschluss über den wesentlichen Inhalt eines Textes geben.

Probleme und Risiken: Die genannten Verfahren bringen einige Probleme und Risiken mit sich. So muss für die Verwendung der Lemmatisation und des Part-of-Speech Taggings zunächst eine geeignete Lookup-Table vorhanden sein. Weiter sind beide Verfahren anfällig für Rechtschreibfehler. Ein Stemming-Algorithmus hingegen kann ein Wort möglicherweise bis auf einen Buchstaben kürzen und damit unkenntlich machen [16, S. 32].

B. Nachrichten in sozialen Medien

Zum Thema Nachrichten und soziale Medien existieren viele Forschungsarbeiten. Eine Recherche auf den gängigen Plattformen vermittelt den subjektiven Eindruck, dass dabei vorwiegend Twitter als Datenquelle Verwendung findet. Dies ist bei den hier vorgestellten Arbeiten ausschließlich der Fall. Die hier vorgestellten Arbeiten können mögliche Bestandteile bei der Erstellung eines Dossiers sein. Die Nutzung von Twitter als Datenquelle bringt einige Vor- und Nachteile mit sich.

a) Vorteile: Jeden Tag erstellen Nutzer auf Twitter über 500 Millionen Tweets [18] mit Informationen zu den aktuellsten Themen [11] [19]. Dies macht Twitter als Datenquelle attraktiv.

Die vorgegebene maximale Länge von 140 Zeichen zwingt Twitter-Nutzer dazu, sich kurz zu fassen und nur die wesentlichsten Informationen einzubauen [19].

b) Nachteile und Schwierigkeiten: Twitter und soziale Medien im allgemeinen sind stark durch die Englische Sprache geprägt. Der Großteil der Tweets sind in Englisch verfasst. 2013 belief sich der Anteil der in Deutsch verfassten Tweets lediglich auf 0,35%. Dagegen steht ein 51,02-prozentiger Anteil an Englischen Tweets. Ausgehend von 500 Millionen Tweets am Tag [18] beläuft sich die totale Anzahl Deutscher Tweets pro Tag im Mittel auf 1,75 Millionen. Mit dieser Anzahl kann gearbeitet werden. Im Hinblick auf die umgerechnet 255,1 Millionen Englischen Tweets pro Tag ist eine Verwendung letzterer jedoch deutlich attraktiver.

Trotz der großen Menge an Tweets zu einem Thema ist der tatsächliche Informationsgehalt innerhalb dieser Menge relativ gering. Es gibt eine große Redundanz einiger Informationen und in der Ergebnismenge einer Suchanfrage einige nichtrelevante Tweets, die es herauszufiltern gilt [11].

Ein weiteres, oft genanntes Problem bei der Verarbeitung von Tweets ist der spezielle Sprachgebrauch. Durch die auf 140 Zeichen beschränkte Länge von Tweets gebrauchen viele Nutzer Abkürzungen. Zudem ist die Rechtschreibung in Tweets häufig mangelhaft [11] [12]. Ein weiteres Phänomen ist die Internet- und Twitter-spezifische Sprache [12].

1) Zusammenfassung von Events mit Topic Models: Chua und Asur [11] stellen in ihrem Forschungsbericht ein Verfahren zur Zusammenfassung von Events anhand von Tweets vor. Ihr Ansatz ist es, die über eine Stichwortsuche per Twitter-API erhaltene Menge von Tweets auf verschiedene, häufig vorkommende *Topics* hin zu untersuchen. Diese Topics stellen die einzelnen Aspekte des Events dar. Eine Sammlung von Tweets, in der jeder Tweet eine der ermittelten Topics repräsentiert, stellt am Ende die Zusammenfassung dar.

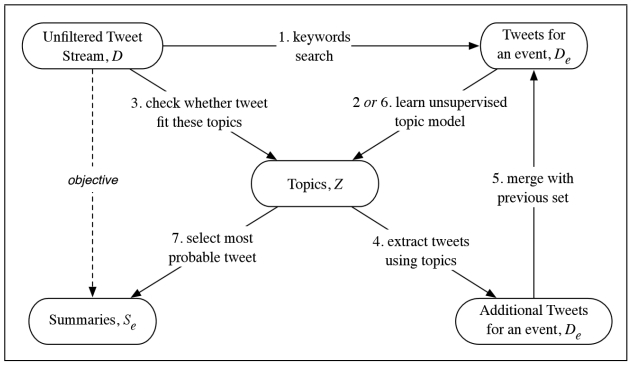


Figure 2. Vorgehensweise bei der Erstellung einer Zusammenfassung [11, S. 82]

Im ersten Schritt beziehen Chua und Asur [11] per dem Ereignis e entsprechender Stichwortsuche eine Menge D_e^1 von Tweets aus der ungefilterten Menge an Tweets D aus einem bestimmten Zeitraum. Aus dieser Menge ermitteln sie im nächsten Schritt eine Menge Z der relevantesten Begriffe (Topics) für das Event e. Die Anzahl der so zu ermittelten Topics ist dabei anpassbar. Die Topics dienen für eine weitere Stichwortsuche in D, um bei der ersten Suche möglicherweise unentdeckte, aber relevante Tweets D_e^2 zu ermitteln. Die beiden Mengen D_e^1 und D_e^2 bilden nach Vereinigung die Gesamtmenge D_e an relevanten Tweets für das Ereignis. Aus diesen Tweets ermitteln die Autoren wiederum die relevantesten Topics, dies aufgrund der größeren Menge mit einer höheren

Genauigkeit als beim ersten Mal. Die nun erlangte Menge an Topics Z stellt die wichtigsten Aspekte des Ereignisses dar. Zuletzt suchen die Autoren für jede Topic $z \in Z$ nach einem repräsentativen Tweet und erhalten somit eine Menge an Tweets, welche die Zusammenfassung S_e des Ereignisses bilden.

Weiter beschreiben die Autoren ein Modell zur Kompensation der Kürze der Tweets, das *Decay Topic Model (DTM)*. Nach diesem Modell erbt ein Tweet den Inhalt seiner direkten Vorgänger, gewichtet nach zeitlichem Abstand zwischen den Tweets. So "verfällt" die Relevanz einer Topic mit der Zeit. Diesem Vorgenen zugrunde liegt die Annahme, dass zeitlich nahe beieinanderliegende Tweets zu einem Thema über einen ähnlichen Inhalt verfügen. Durch Anwendung dieser Methode erhalten die Autoren einen zeitlichen Verlauf der einzelnen Aspekte eines Ereignisses. Dieser ist in 3 dargestellt.

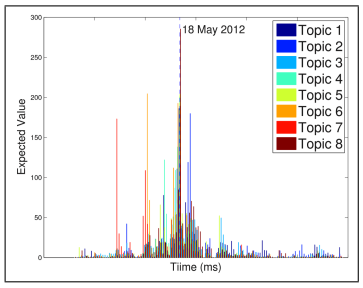
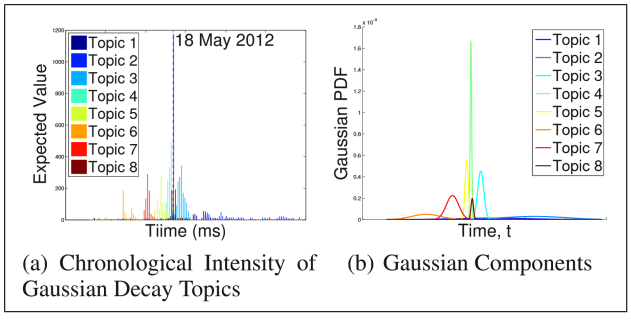


Figure 3. Zeitlicher Verlauf der Topics eines Ereignisses [11, S. 85]

Die Grafik gibt grob Aufschluss über den Verlauf, zeigt aber auch, dass das DTM die Topics nicht ausreichend differenziert. Daher erweitern die Autoren ihr Modell um einen Gauß-Faktor zum *Gaussian Decay Topic Model (GDTM)*. Hierfür nehmen sie an, dass die Relevanz eines Aspektes zeitlich Gauß-verteilt ist. Abbildung 4 zeigt das Ergebnis bei Anwendung des Gauß-Faktors.



(a) Chronological Intensity of Gaussian Decay Topics (b) Gaussian Components

Figure 4. Zeitlicher Verlauf der Topics eines Ereignisses mit Gauß-Faktor [11, S. 87]

Es ist ein klarer zeitlicher Verlauf der Relevanz der einzelnen Topics zu erkennen. Als weitere Schritte nennen die Autoren personalisierte Zusammenfassungen und die Möglichkeit, sachliche Tweets stärker zu gewichten und somit eine höhere Qualität der Zusammenfassung zu erlangen.

2) Zusammenfassung von Events unter Verwendung von Sprachhandlungen: Zhang et al. [12] beschreiben in ihrer Arbeit den Ansatz, Ereignisse unter Verwendung von Sprachhandlungen zusammenzufassen. Hierbei wird jedem Tweet in einer Menge aus für ein Ereignis relevanten Tweets eine Sprachhandlung zugewiesen, was folgend die Extraktion der enthaltenen Information erleichtert und zusätzlich beispielsweise zwischen Fakten und persönlichen Meinungen unterscheiden lässt.

Zunächst beschreiben die Autoren, dass sie für die Auswertung der Tweets eigene Lexika schreiben mussten, welche sie unter anderem für die Identifizierung der Sprachhandlungen verwenden [12]. Dies ist aufgrund der erwähnten Twitter-eigenen Sprache notwendig.

Als die vier möglichen zu identifizierenden Sprachhandlungen nennen die Autoren *statement* (Aussage), *question* (Frage), *suggestion* (Anregung) und *comment* (Kommentar). Weiter führen sie die Kategorie *miscellaneous* (Sonstige) ein. Diese Handlungen lassen sich über Schlüsselbegriffe oder -Phrasen identifizieren. Alternativ können Satzzeichen und andere Symbole auf eine bestimmte Sprachhandlung hinweisen. Die Twitter-spezifischen Symbole *#*, *@* und *RT* können auf zwischenmenschliche Interaktion hinweisen.

TABLE III
DETAILS OF EXPERIMENTAL DATASETS

Category	Topic	# Tweets
News	Japan Earthquake	1742
	Libya Releases	1408
Entity	Dallas Lovato	677
	Nikki Taylor	786
LST	#100factsaboutme	2000
	#sincewebeinghonest	2000

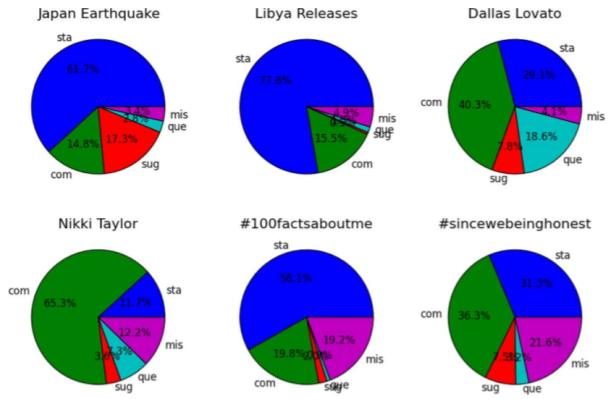


Fig. 2. Speech Act Distributions in the 6 Twitter Topics.

Figure 5. Beispiele für Themen mit Verteilung der Sprachhandlungen [12, S. 652]

Abbildung 5 zeigt sechs Beispielthemen und die Verteilung der Sprachhandlungen in den zugehörigen Tweets. Es ist offensichtlich, dass bei Nachrichtenthemen die Aussagen dominieren, während bei Interessensobjekten (Entity) wie Per-

sonen die Kommentare überwiegen. Bei den *Long-standing Topics (LST)* lässt sich keine klare Aussage treffen, da hier kein einheitlicher Typ vorliegt.

Nach der Ermittlung der Sprachhandlung können die Hauptsprachhandlungen zu einem Thema ermittelt werden. Diese definieren die Autoren als solche, welche mindestens 20% der Menge an Tweets ausmachen [12]. Die nun vorliegende Menge wird weiterverarbeitet, indem zunächst ein Part-of-Speech Tagging durchgeführt wird. Das Wissen über die Sprachhandlung und das Tagging ermöglichen die Extraktion der wesentlichen Informationen. So sind in Aussagen und Kommentaren die Nomen und Nomenphrasen die entscheidenden Begriffe, wobei in einem Kommentar ein Meinungswort vorkommen muss. In einer Anregung hingegen sind Verben und Verbphrasen die Schlüsselbegriffe.

Die so erlangten Schlüsselwörter und -Begriffe fasst ein Algorithmus unter Verwendung einer Schablone zu einer Zusammenfassung zusammen. Ein Beispiel einer solchen Zusammenfassung ist in Abbildung 6 neben den Ergebnissen anderer Methoden dargestellt.

TABLE IX
HUMAN AND AUTOMATIC SUMMARIES FOR #*AGOODBOYFRIEND*

Human	*People are tweeting the qualities that make a good boyfriend and the things a good boyfriend does.*
Our method	*For "a good boyfriend", people state "Team Minaj, DAMN Derrick Rose, Yuri Gagarin" and comment on "love joy, silent cries, good girlfriend".*
SumBasic	*#agoodboyfriend is #agoodboyfriend whether he's around u or not.. "#AGoodBoyfriend" is really a TT ? #agoodboyfriend is not looking for #ago*
Hybrid TF-IDF	*RT @DamnItsTrue: GREAT LIFE = Good Friends Good Food Good Song #agoodboyfriend #DamnItsTrue @DamnItsTrue: GREAT LIFE = Good Friends +*

Figure 6. Beispiele für eine Zusammenfassung [12, S. 656]

Aus dem Satz (*Our method*) lässt sich erkennen, dass Aussagen (*state*) und Kommentare (*comment on*) die ermittelten Hauptsprachhandlungen sind. Für weitere Arbeiten nehmen sich die Autoren vor, unterschiedliche Klassifikatoren auszutesten und die Ergebnisse zu vergleichen. Weiter möchten sie die Lesbarkeit der Zusammenfassung verbessern

3) Erkennung von Nachrichtenverwaltern: Ein möglicher und interessanter Schritt in der Verbesserung der Nachrichtenerkennung auf Twitter ist es, sogenannte Nachrichtenverwalter (*News Curators*) zu identifizieren. Diese Nutzer sind menschlich und auf möglichst wenige Themengebiete spezialisiert. Solche Nutzer sind potentielle Quellen für neue Aspekte zu einem bekannten Nachrichtenthema.

Lehmann et al. [13] beschreiben in ihrer Arbeit ihr Vorgehen, um solche Nutzer auf Twitter zu identifizieren. Zunächst ermitteln sie die Gruppe an Nutzern, welche die URL eines bestimmten Nachrichtenartikels geteilt haben. Diese Nutzer sind potentielle Nachrichtenverwalter. Allerdings existieren viele automatische Twitter-Accounts, welche lediglich die URL des Artikels, jedoch keine weiteren Informationen geteilt haben. Diese Accounts müssen erkannt und aus der Gruppe entfernt werden. In der Regel lassen sich solche Accounts daran erkennen, dass sie fast ausschließlich URLs ohne weiteren Text (abgesehen von der Überschrift des Artikels) teilen [13]. Außerdem können Accounts, welche extrem häufig Tweets erstellen, als automatisiert angesehen werden. Durch Ausschluss dieser Nutzer erhalten die Autoren eine Untergruppe, welche mit hoher Wahrscheinlichkeit nur noch aus menschlichen Nutzern besteht.

Aus diesen menschlichen Nutzern gilt es nun diejenigen herauszufiltern, welche auf möglichst wenige Themengebiete spezialisiert sind. Hierfür beschreiben die selben Autoren in einer anderen Arbeit [14] eine Methode, mit der sich herausfinden lässt, ob ein Nutzer bereits weitere Artikel geteilt hat, welche für den ursprünglichen Artikel thematisch relevant sind. Sämtliche Nutzer, welche nicht mindestens einen solchen weiteren Artikel geteilt haben, schließen die Autoren als nicht ausreichend am Thema interessiert aus.

Weiter kann ermittelt werden, aus welchen Kategorien einer Nachrichtenseite ein Nutzer Artikel teilt. Die Anzahl dieser Kategorien gibt ebenfalls Aufschluss darüber, wie sehr ein Nutzer am Thema des Artikels interessiert ist. Durch Entfernung derjenigen Nutzer, welche zu viele unterschiedliche Themen bedienen, erhalten die Autoren eine Liste von Nutzern, welche Menschlich und am Thema des ursprünglichen Artikels interessiert sind.

Ein weiteres Kriterium ist die “Sichtbarkeit” eines Nutzers. Jeder Nutzer mit weniger als 1000 Followern wird ausgeschlossen. Von den so übriggebliebenen Nutzern erhoffen sich die Autoren weitere Informationen und eventuell auch Expertenwissen.

C. Dossiererstellung

Die Erstellung von Dossiers ist ein spezieller Anwendungsfall für Text Mining. Hierbei geht es darum, zu einem Thema möglichst die passendsten Inhalte zu präsentieren. Dabei kann es eine zeitliche, geographische oder andere Einschränkungen geben. Weiter kann sich ein Dossier nur auf Nachrichtenartikel beschränken, genauso aber auch aus einer multimedialen Mischung bestehen.

Es gibt verschiedene Möglichkeiten, ein Dossier zusammenzustellen. Schöneberg beschreibt in seiner Arbeit [7] beispielsweise den Weg, einen Artikel als Referenz einzulesen und durch Anwendung von Distanzfunktionen thematisch ähnliche Artikel zu ermitteln. Hierfür beschränkt er die Definition des Begriffs Dossier auf die Zusammenstellung ähnlicher Artikel. Als Quelle verwendet er das Archiv des Eurozine Netzwerks, welches unter www.eurozine.com erreichbar ist.

Im wesentlichen ist die Zusammenstellung eines Dossiers mit einem *Recommender System* vergleichbar. Basierend auf Wissen über Präferenzen sucht ein solches System diesen Präferenzen entsprechende Produkte. Für ein Dossier bilden das Hauptthema und gegebenenfalls die Einschränkungen die Präferenzen und statt nach Produkten sucht das System nach Artikeln und anderen Medieninhalten.

Bancu et al. [20] beschreiben beispielsweise zwei Hauptansätze zur Empfehlung von Nachrichtenartikeln. Ein System

kann basierend auf bekannten Präferenzen und unter Berücksichtigung des Inhalts eines Artikels arbeiten oder aber anhand der Präferenzen anderer Nutzer versuchen, die Eignung eines Artikels vorherzusagen. Hier wird davon ausgegangen, dass Nutzer, welche bisher ähnliche Interessen hatten, am selben Artikel interessiert sein könnten.

Weiter stellt die Arbeit ein Entwickeltes Recommender System für Nachrichtenartikel vor. Allgemein finden sich eher Forschungsarbeiten zu Recommender Systemen als zur Dossiererstellung. Dies ist insofern kritisch, dass Recommender Systeme meist persönliche Präferenzen berücksichtigen, eine Dossiererstellung dies aber nicht notwendigerweise tut. Dennoch lassen sich die Algorithmen übertragen.

Park et al. [21] stellen in ihrer Arbeit einen Ansatz zur Umgehung der Tendenziösität in der Berichterstattung vor. Ihr System empfiehlt dem Leser eines Artikels weitere Artikel zum selben Thema, welche dieses allerdings aus einer anderen Perspektive beleuchten. So soll die selbstständige Meinungsbildung des Lesers gefördert werden.

IV. Fazit und Ausblick

Diese Arbeit gibt eine Einführung in den Begriff des Text Mining. Dieses ist ein wesentlicher Bestandteil bei der Generierung von Dossiers. Die hier vorgestellten Arbeiten zeigen, dass es einige interessante Ansätze gibt, speziell aus Twitter aktuelle Informationen zu beziehen. Die Arbeiten zeigen ebenfalls, dass es möglich ist, via Twitter zu einem bestimmten Thema oder Ereignis gezielt die wichtigsten Informationen zu erlangen. Auch eine Ermittlung weiterer Informationen ist möglich.

In Zukunft wird sich der Autor weiter eingehend mit der hier präsentierten Thematik befassen. Besonders gilt es, die vorgestellten Arbeiten auf ihre Eignung für die Verwendung bei einer Dossier-Generierung hin zu überprüfen. So wäre es beispielsweise denkbar, die mit der in Kapitel III-B1 vorgestellten Methode erstellte Zusammenfassung als zentralen oder ergänzenden Bestandteil eines Dossiers zu verwenden. Weiter wäre es möglich, bekannte Nachrichtenverwalter (Kapitel III-B3) als ergänzende Quelle für ein Dossier zu einem Thema zu verwenden. Ebenfalls ist es denkbar, eine Überprüfung durchzuführen, inwieweit die Verwendung dieser Nachrichtenverwalter als einzige Quelle für ein Themendossier brauchbare Ergebnisse liefert.

Einen für die zukünftige Arbeit besonders interessanten Ansatz stellt Hälker in ihrer Arbeit [3] vor. Die Arbeit von Ritter et al. [19] beschreibt ebenfalls die Ereigniserkennung auf Twitter und darüber hinaus die Generierung eines Ereigniskalenders auf Grundlage der ermittelten Daten. Der wesentliche Unterschied zu den vorgestellten Arbeiten liegt darin, dass die Erkennung ohne vorgegebenes Thema erfolgt. Zudem ist hier der Quellcode frei verfügbar. Unter statuscalendar.com kann das Ergebnis der Arbeit betrachtet werden. Ein möglicher Ansatz, diese Arbeit zu verwenden wäre es, die Kalenderdarstellung um Dossiers zu den ermittelten Ereignissen zu erweitern, sodass ein Kalender mit Dossiers zu den relevantesten Themen auf Twitter entsteht. Zudem besteht die Möglichkeit, das vorhandene System eingehend zu analysieren und eventuell Verbesserungen vorzunehmen.

References

[1] H. Hippner and R. Rentzmann, "Text mining," Informatik-Spektrum, vol. 29, no. 4, 2006, pp. 287–290. [Online]. Available: http://dx.doi.org/10.1007/s00287-006-0091-y

[2] J. H. Kroeze, M. C. Matthee, and T. J. D. Bothma, "Differentiating data- and text-mining terminology," in Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology, ser. SAICSIT '03. Republic of South Africa: South African Institute for Computer Scientists and Information Technologists, 2003, pp. 93–101. [Online]. Available: http://dl.acm.org/citation.cfm?id=954014.954024

[3] N. Hälker, "Text Mining für Newssites," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2014.

[4] ——, "Information Retrieval - Grundlage für Journalismus im Web 2.0," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2014.

[5] M. Schöneberg, "Weak Signals," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2013.

[6] ——, "Ansätze zur Trenderkennung in Texten," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2014.

[7] ——, "Automatisierte Erstellung von Pressedossiers durch Textmining," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2015.

[8] N. With, "Vergleich von Text Mining Algorithmen in Social Media in Bezug auf ein Epidemie-Frühwarnsystem," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2013.

[9] I. Demin, "second screen - next level experience," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2014.

[10] ——, "Text Mining for Second Screen," Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung, 2014.

[11] F. C. T. Chua and S. Asur, "Automatic summarization of events from social media," in ICWSM. Citeseer, 2013.

[12] R. Zhang, W. Li, D. Gao, and Y. Ouyang, "Automatic twitter topic summarization with speech acts," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 3, March 2013, pp. 649–658.

[13] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, "Finding news curators in twitter," in Proceedings of the 22Nd International Conference on World Wide Web Companion, ser. WWW '13 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 863–870. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487788.2488068

[14] ——, "Transient news crowds in social media." in ICWSM, 2013.

[15] P. Buneman, "Semistructured data," in Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ser. PODS '97. New York, NY, USA: ACM, 1997, pp. 117–121. [Online]. Available: http://doi.acm.org/10.1145/263661.263675

[16] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval. Cambridge university press Cambridge, 2008, vol. 1.

[17] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," Journal of information science, vol. 18, no. 1, 1992, pp. 45–55.

[18] Twitter, "About Twitter, Inc." Zugriff am 28. Februar 2015, verfügbar unter about.twitter.com/company.

[19] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 1104–1112. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339704

[20] C. Bancu, M. Dagadita, M. Dascalu, C. Dobre, S. Trausan-Matu, and A. M. Florea, "Arsys–article recommender system," in Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2012 14th International Symposium on. IEEE, 2012, pp. 349–355.

[21] S. Park, S. Kang, S. Chung, and J. Song, "Newscube: delivering multiple aspects of news to mitigate media bias," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2009, pp. 443–452.