Ein Ansatz zu Opinion Mining und Themenverfolgung für eine Medienresonanzanalyse

Thomas Scholz

Heinrich-Heine-Universität Düsseldorf
Institut für Informatik
Universitätsstr. 1
D-40225 Düsseldorf, Deutschland
scholz@cs.uni-duesseldorf.de

pressrelations GmbH
Entwicklung
Klosterstr. 112
D-40211 Düsseldorf, Deutschland
thomas.scholz@pressrelations.de

Zusammenfassung

Heutzutage gibt es eine unüberschaubare Anzahl von Medien mit enorm vielen Artikeln und Beiträgen. Da in ihnen neben vielen anderen potenziell nützlichen Informationen wertvolle Meinungen zu Themen enthalten sind, ist eine automatische Beobachtung dieser Medien sehr interessant, birgt aber zwei große Herausforderungen: eine automatisierte Tonalitätsbestimmung (Opinion Mining) kombiniert mit einer Themenverfolgung. Diese zwei Aufgaben sind Teilgebiete des Text Minings, auch Text Data Mining oder Knowledge Discovery in Texten genannt. Diese Arbeit beschreibt einen Ansatz für Opinion Mining und Themenverfolgung basierend auf einer Information Extraction Architektur. In der Evaluation wird gezeigt, wie dieser Ansatz für Opinion Mining oder eine Themenverfolgung eingesetzt werden kann.

Kategorie

Data Mining and Knowledge Discovery

Schlüsselwörter

Opinion Mining, Topic Tracking, Text Mining

1. EINLEITENDE MOTIVATION

Das Internet, Printmedien, TV, Hörfunk und Soziale Netzwerke sind eine Fundgrube von Meinungen zu bestimm-ten Themen in Form von Artikeln oder Beiträgen. Heutzutage ist es möglich diese in digitaler Form zu erfassen. Im Online Bereich können beispielsweise Crawler eingesetzt werden um die Artikelseiten von Internetnachrichten zu erfassen. Durch Analyse des Seitenquelltextes und den Einsatz von Heuristiken kann der eigentliche Artikeltext gefunden werden. Artikel aus Printmedien können eingescannt und mit optischer Zeichenerkennung (OCR) digitalisiert werden. Artikel und Beiträge in digitaler Textform bieten auch teilweise die TV- und Hörfunksender an. Im Bereich Soziale Netzwerke haben sich Dienstleister darauf spezialisiert, die Diskussionen und Kommentare aus Netzen wie Twitter, Facebook oder bestimmte Foren automatisch zu erfassen und als Daten anzubieten.

Die Auswertung von einer großen Menge dieser Artikel und Beiträge ist hingegen schwierig. Eine manuelle Auswertung ist nur mit erheblichen Aufwand möglich und für eine automatische Auswertung gibt es erste Ansätze, aber längst noch keine allumfassende Lösung. Mit der Aufgabe diese Daten zu erfassen und auszuwerten beschäftigt man sich beim Medienmonitoring bzw. bei der Medienresonanzanalyse. In diesem Bereich arbeiten Ausschnittdienste.

1.1 Medienmonitoring

Die riesigen Ströme aus Artikeln und Beiträgen enthalten viele potenziell wertvolle Informationen zu Themen, Personen, Firmen, Produkten, usw. Besonders die PR- und Marketing-Abteilungen von Unternehmen, Parteien und Verbänden interessieren für diese Daten und deren Auswertung. Dabei interessiert man sich besonders dafür, inwiefern sich das Image des Unternehmens oder das Image von bestimmten Produkten, Marken und Dienstleistungen entwickelt. Aber auch wie bestimmte Personen (Werbeträger, Vorstandsmitglieder, etc.) in diesen Medien wahrgenommen werden. Außerdem von Bedeutung ist die Frage, auf welche Weise bestimmte Themen mit dem Unternehmen verknüpft sind. Beim Medienmoitoring geht es darum Artikel zu erfassen und sammeln, die für PR- und Marketing Abteilungen interessant sind. Dazu werden von ihnen Themen oder Schlagwörter definiert. Dies ist auch interessant für Verbände, Vereine, Parteien, Stiftungen, die teilweise zu klein sind um über eine eigene PR-Abteilung zu verfügen.

1.2 Medienresonanzanalyse

Bei der Medienresonanz geht es darum zu bestimmten Themen das mediale Echo zu analysieren.

Dies kann z. B. auf folgende Art und Weise geschehen: Zunächst werden Themen definiert, die es zu untersuchen gilt. Dies können beispielsweise Marken von Firmen sein oder andere Begriffe wie Produktnamen, Personen oder ähnliches. Bei einem Medienbeobachter und einem Ausschnittdienst würden die Kunden (meist PR-Abteilungen von Firmen oder Organisationen wie Parteien) Themen durch bestimmte Schlagwörter festlegen. Die Schlagwörter werden dann von Crawlern in den Medien gesucht, um die entspre-

23^{rd} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 31.05.2011 - 03.06.2011, Obergurgl, Austria.

chenden Artikel zu erfassen.

Dann können diese Artikel bewertet werden, z.B. ob sie wirklich relevant zu diesem Thema sind, wie exklusiv dieser Beitrag ist und am wichtigsten welche Tonalität er besitzt. Die Tonalität beschreibt, ob ein bestimmter Artikel positive oder negative Meinungen zu einem Thema enthält. Es ist auch beides möglich. Oft wird dies mit einem negativen Zahlenwert für eine negative Tonalität und einem positiven Zahlenwert für eine positive Tonalität festgehalten. Diese Bestimmung der Tonalität wird heutzutage bei den Medienbeobachtern noch rein manuell durch Medienanalysten ausgeführt, die die Texte lesen und meinungsbeinhaltende Passagen identifizieren und bewerten.

Allerdings stoßen solche Beobachtungsdienste aufgrund der Menge der Artikel und Beiträge heute an ihre Grenzen, was manuell noch zu bearbeiten ist. Die Nachrichtenanzahl, die diese Dienste bearbeiten, nimmt stetig zu, aufgrund der stärkeren Digitalisierung der Medien wie auch durch Entstehen neuer Medien im Internet wie der Bereich Social Media.

1.3 Opinion Mining und Themenverfolgung als Lösung

Als Lösung für eine automatische Medienresonanzanalyse bieten sich Opinion Mining und Themenverfolgung an, um die doch bisher meist manuell geleistete Arbeit zu unterstützen und perspektivisch zu ersetzen.. Mit Hilfe des Opinion Minings soll es gelingen meinungstragende Passagen innerhalb eines Textes zu finden und dann automatisch mit einem Tonalitätswert zu versehen. Durch die Realisierung einer automatischen Themenverfolgung könnte die manuell vorgenommene Schlagwortverwaltung für die Themenzuordnung abgelöst werden. In der Kombination hätte man dann eine automatische Medienresonanzanalyse.

Der Rest dieser Arbeit kann wie folgt zusammengefasst werden: Zunächst werden die Ansätze aufgezeigt, die schon zum Bereich des Opinion Minings und der Themenverfolgung entwickelt wurden. Dann wird die Architektur entworfen, wodurch eine automatische Medienresonanzanalyse auf Basis von Natural Language Processing und Information Extraction realisierbar wird. Anschließend wird in ersten Versuchen demonstriert, wie die Vorverarbeitung für Opinion Mining und Themenverfolgung eingesetzt werden können. Abschließend werden dann aufgrund dieser Ergebnisse Schlussfolgerungen für das weitere Vorgehen gezogen.

2. VERWANDTE ARBEITEN

2.1 Opinion Mining

Um die Tonalität eines Textes zu bestimmen, benutzen viele Ansätze wie [2, 3, 9, 13] Wörterbücher, in denen Wörter mit einem Tonalitätswert hinterlegt sind. Diese Wörterbücher werden meist so aufgebaut: Man beginnt mit einer kleineren Menge von positiven und negativen Wörtern. Dann wird analysiert, ob neue Wörter oft mit positiven oder negativen Wörtern auftauchen und entsprechend bekommt dann das neue Wort einen positiven oder negativen Tonalitätswert.

Eine typische Menge von Saattonalitätswörtern sieht beispielsweise so aus [16]:

- positiv: {good, nice, excellent, positive, fortunate, correct, superior}
- negativ: {bad, nasty, poor, negative, unfortunate, wrong, inferior}

Diesen Aufbau durch einen Bootstrapping Algorithmus benutzt auch das bekannteste Wörterbuch: SentiWordNET [5]. Als Quelle für neue Wörter benutzt es dazu Glossetexte und Wortbeziehungen in WordNET, dem bekanntesten englischen, digitalen Wörterbuch. Mit diesen Wortbeziehungen wie Synonym, Oberbegriff oder Unterbegriff werden neue Wörter gefunden, die dann die gleiche oder eine ähnliche Tonalität bekommen.

Andere Ansätze benutzen nur WordNET oder ähnliche, allgemeine Wörterbücher [6, 7] oder Textsammlungen [2, 4] oder Suchanfragen [6]. Die Vorgehensweise ähnelt dabei oft der Vergrößerung der Wortmenge durch Bootstrapping.

Viele Ansätze [2, 3, 4, 13] beschränken sich auf das Gebiet der Sentiment Analysis, also der Tonalitätsbestimmung in Kundenrezensionen. Einige Ansätze beschränken sich dabei nur auf Adjektive [3].

Bei Kundenrezensionen ist eine Identifikation von Meinungsblöcken nicht nötig. Eine Rezension besteht nur aus Meinungsblöcken. Um generell Meinungen zu finden, auch wenn dies in einem langen Zeitungsartikel nur ein kleiner Absatz über ein bestimmtes Unternehmen ist, sind zunächst noch anderere Schritte zuvor nötig. Ein satzbasierter Ansatz [7] bestimmt für jeden Satz einen Tonalitätswert basierend auf den in ihm enthalten Wörtern. Dafür werden in zwei Modellen einmal alle Wörter oder nur das stärkste Tonalitätswort herangezogen. Überschreitet der Wert eine gewisse Grenze, dann enthält der Satz eine Meinung. Für einen anderen Ansatz [1] enthält ein Satz eine Meinung, wenn er ein Adjektiv enthält.

Auch kann man davon ausgehen, dass bei Sentiment Analysis Ansätzen man schon mit einem kleineren Wörterbuch mit Tonalitätswerten zurecht kommt, da man es bei den Zielen der Rezension nur mit Produkten und ähnlichem wie Filmen, Hotels usw. zu tun hat. Bei einer Medienresonanzanalyse werden aber gleichzeitig Entitäten wie Personen, Organisationen, Produkte, Events oder Aktionen im Fokus stehen. Somit ändern sich auch die tonalitätsbildenen Wörter, da nicht allein durch eine Beschreibung eines Produktes etc. eine Tonalität ausgedrückt wird.

In einer Rezension will der Autor seine Meinung dem Leser direkt vermitteln. In Zeitungsartikeln beschreibt der Autor nicht nur direkte Meinungen, oft wird eher über Fakten und Handlungen gesprochen, die sich auf bestimmte Personen oder Organisationen beziehen, die dann eine Tonalität entstehen lassen. Darum sollte ein solcher Ansatz auch nicht nur Adjektive, sondern mehr Wortarten miteinbeziehen (z. B. Verben). Die meisten Ansätze [1, 2, 3] sind auf die englische Sprache ausgerichtet. Darüber hinaus gibt es einige Ansätze für die Chinesische Sprache, die allerdings nicht die Güte der Ergebnisse der auf Englisch arbeitenden Ansätze erreichen [4, 9].

2.2 Themenverfolgung

Wissenschaftliche Methoden [10, 14, 17, 18], die eine Themenverfolgung realisieren, stellen ein Thema oft durch Schlagwörter dar. Diese Schlüsselwörter werden dadurch extrahiert, dass die häufigsten Wörter eines Themas genommen werden [14], die TF-IDF Methode zur Gewichtung benutzt wird [10] oder die Wörter ausgewählt werden, die am wahrscheinlichsten in einem Thema vorkommen und am unwahrscheinlichsten in allen anderen Themen [17, 18].

Weiterhin gibt es einen Ansatz [8] einzelne Personen zu verfolgen. In diesem Ansatz geht es dann später um eine Visualisierung der Daten: Wie oft wurde die Person in den beobachteten Medienquellen in einem bestimmten Zeitintervall (beispielsweise an einem Tag) erwähnt.

Eine andere, sehr erfolgreiche Methode ist die Verfolgung von wörtlicher Rede [11] für ein bestimmtes Thema. Die Arbeit beabsichtigt zu erforschen, wie sich Themen zwischen den verschieden Medien (in diesem Fall Onlinenachrichten und Soziale Netzwerke) bewegen. Die Autoren untersuchen nach welcher Zeit Themen in die sozialen Netzwerke gelangen und ob es Themen gibt, die zuerst in den Sozialen Netzen entstehen und dann erst in die herkömmlichen Nachrichten gelangen. Hier werden Zitate aus wörtlicher Rede benutzt, da diese laut den Autoren einfach zu verfolgen sind [11]. Ein Zitat steht dann für ein Thema. Durch einen graphbasierten Ansatz werden Zitate auch wieder erkannt, wenn sie verkürzt oder leicht abgeändert werden.

Selten werden verschiedene Merkmale kombiniert [12], um Themen darzustellen oder zu verfolgen. Allerdings verlangt dies auch größeren Aufwand, da man zunächst mittels Information Extraktion Methoden viele Informationen im Vorlauf erfassen muss, damit man daraus entsprechende Merkmale generieren kann. Hier sind auch begrenzte Rechnerkapazitäten ein nicht zu vernachlässigender Aspekt.

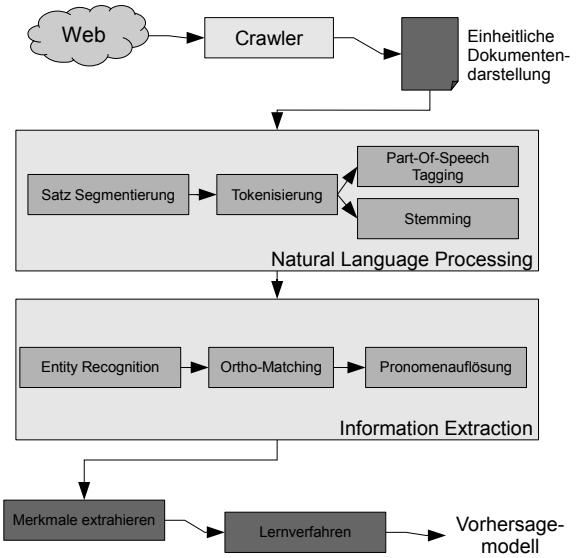

Abbildung 1: Ablauf der Verarbeitungsschritte

3. ANFORDERUNGEN FÜR DEN ANSATZ

Die verschiedenen Arbeiten [1, 2, 3, 7, 12] zu diesen Bereichen zeigen oft, dass es sinnvoll ist, die Texte einer Vorverarbeitung zu unterziehen (s. Abbildung 1). Sehr vorteilhaft erscheint der Einsatz von Natural Language Processing und Information Extraction. Wenn ein Ansatz diese Vorgaben für eine Vorverarbeitung erfüllt, dann entstehen neue Möglichkeiten, die später aufgeführt werden.

3.1 Natural Language Processing

Gerade Natural Language Processing (NLP) wird in vielen anderen Ansätzen benutzt, um unter anderem Adjektive zu identifizieren [1, 2, 3]. Beim Natural Language Processing ist nach dem simplen Aufteilen des Textes in Sätze und Wörter das sogenannte Part-Of-Speech Tagging der wichtigste Analyseschritt. Dabei werden die Wörter aufgrund von Wahrscheinlichkeitsmodellen wie Hidden Markov Modellen grammatikalischen Wortarten wie Nomen, Verben, Adjektiven usw. zugeordnet. Außerdem sollte ein Stemming durchgeführt werden, damit alle Wörter auch in Ihrer Grundform verfügbar sind. Mit dieser Zurückführung werden viele Methoden vereinfacht, die auf der Identifikation von bestimmten Wörtern beruhen oder einen Text als Wortlisten mit Häufigkeiten darstellen.

3.2 Information Extraction

Durch Information Extraction (IE) [15] ist es darüber hinaus möglich, Entitäten im Text wie Personen, Organisationen und Orte zu erkennen. Diese Named Entity Recognition (NER) ist ein gutes Beispiel, wie IE auf NLP aufbaut: Zuerst werden Nomen identifiziert und z.B. durch Listen genauer bestimmt, ob es eine Person ist und eventuell zusätzlich, ob es nur ein Vorname, Vor- und Nachname ist usw. Diese Entitäten werden dann über den Text verfolgbar, wenn weitere Techniken wie Ortho-Matching und eine Pronomenauflösung durchgeführt wird. Ortho-Matching beschreibt das Erkennen der Entität im selben Text an mehreren Stellen, wenn auch nicht exakt die selbe Zeichenkette verwendet wird. Im Falle einer Person könnte z.B. erst der komplette Name und später im Text nur noch der Nachname benutzt werden. Kommt eine Pronomenauflösung hinzu, dann wird eine Entität auch dann weiterverfolgt, wenn im Text für die Entität nur noch Pronomen stehen wie "sie" oder "ihn", die sich aber auf die entsprechende Person beziehen.

Darüber hinaus kann es noch sehr nützlich sein, weitere Informationen wie den grammatikalischen Fall oder das Geschlecht von Wörtern zu bestimmen. Auch Textpassagen mit wörtlicher Rede oder die Extraktion von Nomenphrasen sind weitere nützliche Informationsbausteine, die in einer weiteren Verarbeitung aufgegriffen werden können.

3.3 Neue Möglichkeiten

Diese Vorverarbeitung lässt sich für das Opinion Mining und die Themenverfolgung folgendermaßen einsetzen:

Durch die Bestimmung der verschiedenen Wortarten lassen sich für das Opinion Mining Wörterbücher aus Adjektiven, Verben und Adverben extrahieren. Hier stellt sich noch die Frage, wie man diesen Tonalitätswert bestimmt. Viele Ansätze bilden dazu Maße, die auf das Zusammenauftreten mit positiven bzw. negativen Tonalitätswörtern beruhen. Wenn man viele annotierte Meinungsblöcke besitzt, kann man auch Standardansätze aus dem Information Retrieval wie TF-IDF darauf anwenden. Ebenso kann man Bootstrapping einsetzen.

Auch Nominalphrasen können in dem Wörterbuch aufgenommen werden. Diese können auch für die Darstellung eines Themas von großem Nutzen sein.

Durch die Erkennung von Personen, Organisationen usw. kann man bei der Tonalitätsbestimmung unterscheiden, ob nun eine Person oder ein Produkt besprochen wird. Dadurch kann man die Bewertung des Vokabulars darauf anpassen.

Bei der Themenverfolgung kann man Themen durch die Anwesenheit von Entitäten beschreiben. Das Vorhandensein bzw. die Abwesenheit einer Entität kann ebenso wie Schlagwörter dazu benutzt werden ein Thema zu beschreiben und

damit auch zu verfolgen. Auch könnte man genauso wie bei Schlagwörtern das Auftreten mit TF-IDF gewichten, was nun konkret bedeuten würde: Die Häufigkeit einer Entität multipliziert mit der inversen Dokumentenfrequenz in der sie vorkommt. Bei dieser Gewichtung ist noch zu klären, was den gesamten Dokumentenkorpus darstellt. Dies könnte ein zeitlicher Ausschnitt sein (z. B. ein Monat).

Zusätzlich könnte diese Gewichtung interessante Informationen über die Entwicklung eines Themas liefern, weil es anzeigt welche Entitäten in welchen Themen eine starke Rolle spielen.

Dies kann man zusätzlich mit bisherigen Ansätzen für Themendarstellung durch Schlagwörter sowie die Verfolgung von wörtlicher Rede kombinieren.

Diese Vorverarbeitungsschritte benötigen natürlich auch Rechenzeit. Da allerdings dazu auf einem riesigen Datenfeld (den Texten) nur gelesen werden muss, ist eine Parallelisierung der Vorverarbeitung durchführbar.

4. EVALUATION

Um zu überprüfen, wie dieser Ansatz mit NLP und IE für Opinion Mining und eine Themenverfolgung eingesetzt werden kann, wird evaluiert, wie mit bestimmten Wortarten automatisch eine Tonalität bestimmt werden kann und ob mit Entitäten eine thematische Zuordnung möglich ist. Dafür werden zuvor klassifizierte Daten benutzt.

Diese Evaluation soll erste Hinweise geben, ob es grundsätzlich mit diesen Merkmalen möglich ist, die fundamentalen Bausteine einer Medienresonanzanalyse maschinell durchzuführen: Tonalitätsbestimmung und Themenzuordnung.

Dabei geht es auch weniger um die Bestimmmung des optimalen Lernverfahrens. Der Einfachheit halber wurden dazu im ersten Schritt drei typische Klassifikationsverfahren verwendet. Diese bieten sich an, weil die Daten schon vor der Evaluation mit entsprechenden Klassen versehen sind.

4.1 Tonalitätsbestimmung

Für diesen Test wurden 1600 Nachrichtenmeldungen mit 800 positiven und 800 negativen Meldungen analysiert. Um eine Tonalität zu erhalten, wurden mittels NLP Adjektive, Adverbien und Verben aus dem Text extrahiert und mittels Stemming auf ihre Stammform zurückgeführt. Danach wurden invertierte Listen von diesen Dokumenten erzeugt und die einzelnen Terme mittels TF-IDF gewichtet.

Nach der Erzeugung dieser Attribute wurden die Daten in einer 10-fach-über-Kreuz-Validierung durch drei Klassifikationsverfahren getestet: Support Vector Machine (SVM), Naive-Bayes und k-Nearest-Neigbours mit k=7.

Bei den Resultaten zeigte sich, das diese doch recht naive Methode (es wird beispielsweise nicht betrachtet, ob irgendwelche Negationen anwesend sind) tatsächlich erste brauchbare Hinweise geben kann.

Es zeigte sich, dass die Vermutung, nur Adjektive allein würden die Tonalität bestimmen, nicht zutrifft. Die Gruppe der Verben schneidet schon besser ab. Hier scheint der Unterschied zwischen Sentiment Analysis bezogen auf Kundenrezensionen und Opinion Mining bezogen auf Nachrichten deutlich zu werden. Kundenrezensionen beziehen ihre Tonalität wohl eher durch Adjektive ("die Bildqualität ist super" oder "der Autofokus ist zuverlässig")[1], während in Nachrichten dies nicht unbedingt der Fall ist ("Verbraucherschützer warnen vor der 'Umfrucht'-Falle" oder "Rekordjahr: Audi belohnt Mitarbeiter mit Millionen")[2].

Wortart	Klassifikationsverfahren	Genauigkeit
Adjektive		
	Support Vector Machine	80,81 %
	Naive-Bayes	68,34 %
	k-Nearst-Neighbour	53,60 %
Verben		
	Support Vector Machine	82,07 %
	Naive-Bayes	72,29 %
	k-Nearst-Neighbour	56,05 %
Adverbien		
	Support Vector Machine	75,61 %
	Naive-Bayes	66,08 %
	k-Nearst-Neighbour	53,79 %

Abbildung 2: Tonalitätsbestimmung

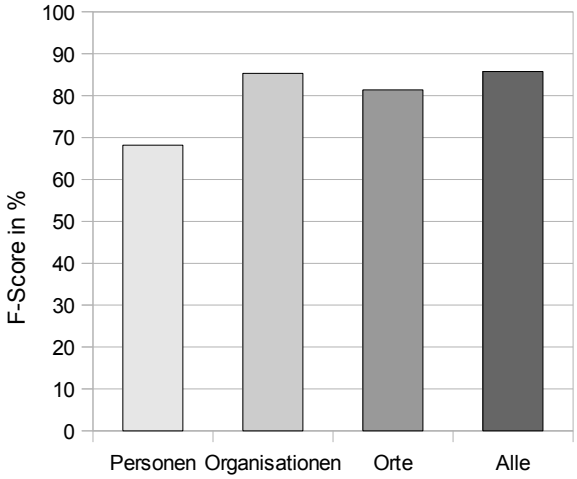

Abbildung 3: Themenzuordnung

Das beste Klassifikationsverfahren der drei hier getesteten Verfahren ist eindeutig die Support Vector Machine. Darum wird sie nun auch im zweiten Teil der Evaluation angewendet.

4.2 Themenzuordnung

Im zweiten Test wurden 204 Texte zu zwei sehr allgemeinen Themen gesammelt. Die Themen waren "Finanzmarkt" und "Parteien".

Dann wurden aus diesen Texten die Entitäten Personen, Organisationen und Orte extrahiert. Diese Entitäten wurden dann wiederum als Terme genommen, die mittels TF-IDF gewichtet wurden. Abschließend wurde eine SVM benutzt um die Daten in einer 10-fach-über-Kreuz-Validierung zu evaluieren.

Die Ergebnisse veranschaulicht Abbildung 3. Interessant ist hier, dass bei den einzelnen Merkmalen die Organisationen das beste Resultat liefern (ca. 85,29 %). Dies kann aber mit der Themenauswahl ("Finanzmarkt" und "Parteien") zu tun haben, in der wahrscheinlich generell mehr Organisatio-

[1]Beispiele aus einer Amazon.de Kundenrezension

[2]Beispiele von Spiegel.de am 4.3.2011

nen eine Rolle spielen bzw. die Organisationen das trennende Kriterium sind.

Auch Orte scheinen charakteristisch für Themen zu sein (ca. 81,37 %). Kaum überraschend spielt der Ort "Frankfurt" im ersten Thema eine wichtigere Rolle als im zweiten Thema und für "Berlin" ist es umgekehrt.

Das Merkmal Person ist nicht so erfolgreich (nur ca. 68,14 %). Bei den hier vorliegenden Themen gab es hinsichtlich der Personen auch durchaus Überschneidungen, weil mehrere Personen in beiden Themen auftauchten. Eine interessante Frage stellt sich nun dahingehend, ob dies bei kleineren Themen vielleicht seltener der Fall ist.

Insgesamt zeigt sich das wünschenswerte Resultat: Mit allen Entitäten gemeinsam wird das beste Ergebnis erzielt (ca. 85,78 %).

5. SCHLUSSFOLGERUNG UND WEITERFÜHRENDE FRAGESTELLUNGEN

Die Ergebnisse der Evaluation lassen darauf schließen, dass sich aufbauend auf dem beschriebenen Anforderungsprofil eine automatische Tonalitätsbestimmung und Themenverfolgung realisieren lässt.

Zu dem Aspekt des Opinion Minings fehlen noch viele Bestandteile, die in einem Text die Tonalität verändern können. Es hat sich gezeigt, dass die Worte allein schon im Ansatz funktionieren, aber noch großes Verbesserungspotenzial vorhanden ist.

Dazu ist zu erarbeiten, ob es noch bessere Methoden der Gewichtung gibt als der Standardansatz über TF-IDF. Außerdem muss überlegt werden, wie man die Tonalitätswörter beispielsweise in einem Wörterbuch verwalten kann. Als nächste Fragestellung schließt sich dann an, wie man mit semantischen Merkmalen wie Negation oder dem Bezug zu Entitäten umgeht.

Darüber hinaus ist ein weiteres spannendes Problem die Identifizierung der Meinungsblöcke, also der Textpassagen, die eine Meinung beinhalten. Ein Tonalitätsgrenzwert für Abschnitte und Sätze ist denkbar, aber auch die Lokalisierung durch die Entitäten im Text, für die man sich erstens verstärkt interessiert und die sich zweitens mit ausreichend vielen tonalitätsbildenen Wörtern umgeben.

Bei der Themenverfolgung haben die Experimente zunächst nur den Wert von Entitätenerkennung in einem einfachen Beispiel gezeigt. Hier müsste die Kombination mit klassischen Schlagwortansätzen und neueren Ansätzen, wie die Einbeziehung von wörtlicher Rede, genutzt werden, um eine bessere Themendarstellung zu erhalten und zusätzlich interessante Fakten über ein Thema zu sammeln. Diese Fakten können Folgendes beinhalten: Wie stark sind welche Personen mit welchen Themen verbunden? Oder gibt es zentrale Zitate/Aussagen, die immer wieder aufgegriffen werden.

Allerdings muss zunächst die Frage beantwortet werden, wie man die Entitäten sinnvoll mit Ansätzen wie Schlagwörtern und die Verfolgung von Zitaten verbinden kann. Dies wird Gegenstand der zukünftigen Arbeit sein, wobei auch zu klären ist, wie man diese Kombination für die Verwaltung einer Themenverfolgung sinnvoll einsetzen kann.

Weiterhin ist dabei die Größe eines Themas zu beachten (für die Defintion der Größe eines Themas gibt es viele Möglichkeiten, die Anzahl der Artikel zu einem Thema ist eine nahe liegende Lösung). Wie wirkt sich die Größe der Themen auf die Verwaltung aus? Und wie verhält sich die Themendarstellung mit Merkmalen dadurch? In der Evaluation kam schon die Frage auf, ob Personen bei kleineren Themen nicht eine wichtigere Rolle zur Themenbeschreibung spielen.

Insgesamt zeigt sich aber, dass die Vorverarbeitung durch Natural Language Processing und Information Extraction von großem Vorteil ist, da sie für beide Aufgabenstellungen, Opinion Mining und Themenverfolgung, viele neue Möglichkeiten eröffnet und diese im Ansatz für eine Medienresonanzanalyse funktionieren.

6. LITERATUR

[1] L. Dey and S. K. M. Haque. Opinion mining from noisy text data. In *Proc. of the 2nd workshop on Analytics for noisy unstructured text data*, AND '08, pages 83–90, 2008.

[2] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proc. of the international conference on Web search and web data mining*, WSDM '08, pages 231–240, 2008.

[3] X. Ding, B. Liu, and L. Zhang. Entity discovery and assignment for opinion mining applications. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1125–1134, 2009.

[4] W. Du and S. Tan. An iterative reinforcement approach for fine-grained opinion mining. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 486–493, 2009.

[5] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proc. of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 617–624, 2005.

[6] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proc. of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 947–956, 2009.

[7] S.-M. Kim and E. Hovy. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005.

[8] M. Krstajic, F. Mansmann, A. Stoffel, M. Atkinson, and D. A. Keim. Processing online news streams for large-scale semantic analysis. In *ICDE Workshops*, pages 215–220, 2010.

[9] L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 100–107, 2006.

[10] S. Lee and H.-J. Kim. News keyword extraction for topic tracking. In *Proc. of the 4th International Conference on Networked Computing and Advanced Information Management - Volume 02*, pages 554–559, 2008.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*,

KDD '09, pages 497–506, 2009.

[12] B. Li, W. Li, Q. Lu, and M. Wu. Profile-based event tracking. In *Proc. of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 631–632, 2005.

[13] M. M. S. Missen, M. Boughanem, and G. Cabanac. Comparing semantic associations in sentences and paragraphs for opinion detection in blogs. In *Proc. of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '09, pages 80:483–80:488, 2009.

[14] X. Tang, C. Yang, and J. Zhou. Stock price forecasting by combining news mining and time series analysis. In *Proc. of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 279–282, 2009.

[15] J. Turmo, A. Ageno, and N. Català. Adaptive information extraction. *ACM Comput. Surv.*, 38, July 2006.

[16] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21:315–346, October 2003.

[17] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 784–793, 2007.

[18] J. Zeng, C. Wu, and W. Wang. Multi-grain hierarchical topic extraction algorithm for text mining. *Expert Syst. Appl.*, 37:3202–3208, April 2010.