Opinion Mining

Nutzung von Twitter als Meinungsquelle zur Vorhersage von Börsenkursen

Masterthesis

zur Erlangung des Grades Master of Science
im Studiengang Wirtschaftsinformatik

vorgelegt von

Peter Valerius

Betreuer:

Dr. Michael Möhring, Institut für Wirtschafts- und Verwaltungsinformatik (IWVI) im FB 4: Informatik der Universität Koblenz-Landau
Prof. Dr. Klaus G. Troitzsch, IWVI, FB 4, Universität Koblenz-Landau

Erstgutachter:

Dr. Michael Möhring, IWVI, FB 4, Universität Koblenz-Landau

Zweitgutachter:

Prof. Dr. Klaus G. Troitzsch, IWVI, FB 4, Universität Koblenz-Landau

Koblenz, im Juni 2012

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Motivation

Die Ausbreitung sozialer Medien im Netz hat in den vergangenen fünf Jahren eine gigantische Datenmenge geschaffen, deren Inhalt fast jeden Teil der Gesellschaft umfasst und ein Ende des Booms ist derzeit noch nicht abzusehen. Durch die zunehmende Offenheit und Mitteilungsbereitschaft der Gesellschaft, bildet die virtuelle Welt des Internets somit immer mehr die reale Welt ab. Dabei sind viele dieser Daten für jeden verfügbar, sehr oft auch ohne das Wissen der Nutzer.

„Today, people have no choice but to give away their personal information—sometimes in exchange for free networking on Twitter or searching on Google, but other times to third-party data-aggregation firms without realizing it at all" (Leber, 2012).

Es existieren zum jetzigen Zeitpunkt allerdings noch sehr wenige Untersuchungen darüber, welche Aussagekraft die im Web verfügbaren Informationen über reale Ereignisse geben können. Als wie repräsentativ und verlässlich können diese Daten angesehen werden und wie entwickelt sich deren Nutzung? Welche Teile der Gesellschaft nutzen soziale Medien? Wie wirkt sich dies auf die Aussagekraft der Daten aus? Hat der Inhalt von Sozialen Medien eine Vorhersagekraft auf reale Ereignisse?

Das Microblogging ist eine der am weitesten verbreiteten Sozialen Medien und somit ein guter Ausgangspunkt für derartige Untersuchungen. Seiten wie Twitter erreichen jeden Tag Millionen von Nachrichten. Die Nutzer reichen dabei von Privatpersonen über Journalisten bis hin zu Politikern und kommen aus allen Teilen der Welt. Als Anbindung nach außen stellt Twitter desweiteren eine API bereit, über die jeder Nachrichten zu allen Themen in Echtzeit abrufen und weiterverarbeiten kann. Die Analyse derartiger Web basierter Daten wird auch als Webmining bezeichnet. Erste Untersuchungen auf diesem Gebiet bescheinigen dem Inhalt von Twitter Nachrichten eine sehr hohe Vorhersagekraft bei den Verkaufszahlen von Kinofilmen. Deutsche Wissenschaftler konnten durch die Analyse von Twitter Nachrichten sehr gute Ergebnisse bei der Vorhersage von Bundeswahlen, beziehungsweise bei der Wahl von Koalitionspartnern treffen. Weitere Analysen bescheinigen Twitter eine hohe Vorhersagekraft von Börsenkursen. Der Londoner Hedgefond Anbieter Derwent Capital Markets nutzt Twitter-Analysen beispielsweise bereits für die Verbesserungen seiner Investmententscheidungen vgl. (Lischka, 2012). Thema der Arbeit ist es diese Quelle weiter zu untersuchen.

Dabei kommt der Untersuchung zugute, dass der Inhalt von Twitter Nachrichten auf 140 Zeichen begrenzt ist, da die maschinelle Analyse und Klassifikation von längeren Fließtexten bisher nur begrenzt nutzbare Ergebnisse liefert. Es lässt sich zwar sehr leicht die Syntax eines Textes untersuchen, also welche Wörter kommen wie oft vor. Die Semantik eines Textes lässt sich hierdurch allerdings nur unzureichend bestimmen. Die geringe Länge von Twitter Nachrichten dagegen schränkt die Interpretationsmöglichkeiten stark ein, was einen Algorithmus, der den Text untersucht, weitaus verlässlicher macht. Eine Form der Textanalyse ist das Opinion Mining. Diese Unterform des Data Mining untersucht die Meinungsausprägung von Texten mit dem Ziel einer Zuordnung von Inhalten in positive,

negative und neutrale Aussagen vorzunehmen. In Verbindung mit den Millionen von Äußerungen auf Twitter ergibt sich hierdurch eine riesige Meinungsbibliothek.

1.2 Zielsetzung

Neben den theoretischen Grundkonzepten der automatisierten Fließtextanalyse, die das Fundament dieser Arbeit bilden, soll ein Überblick in den derzeitigen Forschungsstand bei der Analyse von Twitter-Nachrichten gegeben werden. Hierzu werden verschiedene Forschungsergebnisse der, derzeit verfügbaren wissenschaftlichen Literatur erläutert, miteinander verglichen und kritisch hinterfragt. Deren Ergebnisse und Vorgehensweisen sollen in unsere eigene Forschung mit eingehen, soweit sie sinnvoll erscheinen. Ziel ist es hierbei, den derzeitigen Forschungsstand möglichst gut zu nutzen. Ein weiteres Ziel ist es, dem Leser einen Überblick über verschiedene maschinelle Datenanalysemethoden zur Erkennung von Meinungen zu geben. Dies ist notwendig, um die Bedeutung der im späteren Verlauf der Arbeit eingesetzten Analysemethoden in ihrem wissenschaftlichen Kontext besser verstehen zu können. Da diese Methoden auf verschiedene Arten durchgeführt werden können, werden verschiedene Analysemethoden vorgestellt und miteinander verglichen. Hierdurch soll die Machbarkeit der folgenden Meinungsauswertung bewiesen werden. Um eine hinreichende Genauigkeit bei der folgenden Untersuchung zu gewährleisten, wird auf ein bereits bestehendes und evaluiertes Framework zurückgegriffen. Dieses ist als API[1] verfügbar und wird daher zusätzlich behandelt. Der Kern Inhalt dieser Arbeit wird sich der Analyse von Twitternachrichten mit den Methoden des Opinion Mining widmen. Es soll untersucht werden, ob sich Korrelationen zwischen der Meinungsausprägung von Twitternachrichten und dem Börsenkurs eines Unternehmens finden lassen. Es soll dabei die Stimmungslage der Firma Google Inc. über einen Zeitraum von einem Monat untersucht und die dadurch gefunden Erkenntnisse mit dem Börsenkurs des Unternehmens verglichen werden. Ziel ist es, die Erkenntnisse von (Sprenger & Welpe, 2010) und (Taytal & Komaragiri, 2009) auf diesem Gebiet zu überprüfen und weitere Fragestellungen zu beantworten.

1.3 Aufbau der Arbeit

Nach dieser Einleitung wird im Kapitel zwei gezeigt, wie sich Aktienkurse zusammensetzten und welche Einflussfaktoren auf sie einwirken. Des Weiteren wird auf verschiedene Textanalyseverfahren eingegangen, die als Grundgerüst für alle weiteren Analysen dienen werden. Kapitel drei wird Blogging und Mikroblogging im Allgemeinen und Twitter als Mikroblogging-Dienst im Speziellen beschreiben, sowie dessen Einsatzmöglichkeiten erläutern. Dabei ist zu erwähnen, dass Kapitel zwei und drei analog gelesen werden können, da sie nicht aufeinander aufbauen. Beide bilden jedoch die Grundlage für alle folgenden Kapitel. Kapitel vier geht auf verschiedene Sentiment-Analyseformen ein. Da diese auf verschiedenen Techniken aufbauen können, werden diese einzeln beschrieben und evaluiert. In Kapitel fünf wird die eigentliche Analyse zur Vorhersage von Börsenkursen durchgeführt. Hierin wird der Aufbau des Testsystems beschrieben sowie der Ablauf der Analyse erläutert. Den Schlusspunkt der Arbeit bilden die Evaluation der Forschungsergebnisse und ein Ausblick auf zukünftige Untersuchungen

[1] Ein Application Programming Interface (API) ist ein Teil eines Programmes, der von den Entwicklern als Anbindung an ein Softwaresystem zur Verfügung gestellt wird. Hierdurch können verschiedene Systeme Daten austauschen, ohne die Struktur des jeweils anderen Systems zu kennen.

2 Theoretische Einordnung

2.1 Börsenkurse

Da die Hauptfrage dieser Arbeit die Vorhersagekraft von Twitternachrichten auf Börsenkurse ist, stellt sich die Frage wie sich Börsenkurse zusammensetzen und welche Einflussgrößen auf sie einwirken. Vor allem, welcher logische Zusammenhang überhaupt zwischen öffentlicher Stimmung und einem Aktienkurs besteht. Die nächsten zwei Unterkapitel werden daher dieser Fragestellung nachgehen.

2.1.1 Wie ergeben sich Börsenkurse

Mit dem Kauf einer Aktie erwirbt man einen Anteil am Grundkapital einer Aktiengesellschaft. Der Anteil der Aktie entspricht dem Nennwert der Aktie bezogen auf den Nennwert aller Aktien. Es existieren allerdings auch Aktien ohne Nennwert. Der Anteil am Grundkapital dieser Stückaktien entspricht der Stückzahl an Aktien bezogen auf die Stückzahl aller Aktien. Aktien sind eine Handelsware, die an einem Markt gehandelt werden, dieser Markt ist die Börse. Der Preis einer Aktie resultiert wie bei jeder Handelsware in einer freien Markwirtschaft aus dem Verhältnis von Angebot und Nachfrage. Die Kauf- und Verkaufsentscheidungen aller an den Börsen beteiligten Akteure bestimmen den Kurs. Nach der Kapitalmarktheorie verhalten sich diese Akteure nach dem Prinzip des Homo Ökonomikus. Nach diesem Prinzip versucht jeder Akteur durch den Kauf und Verkauf von Aktien, seinen Gewinn zu maximieren. Da dieser Gewinn jedoch nicht sicher ist, muss jeder Akteur eine Risiko-Nutzenabwägung treffen. Es wird dabei sehr oft unterstellt, dass die Kurse am Aktienmarkt einer Normalverteilung folgen. Geringe Kursschwankungen kommen somit sehr häufig vor, während starke Schwankungen sehr unwahrscheinlich sind.

„Es handelt sich hier um die dritte Abstraktionsstufe des Tausches. Die erste Stufe wäre die Erzeugung von Gütern, ihr direkter Tausch und ihr Verbrauch. Die zweite Stufe wäre der Handel von Gütern auf einem Markt. Die dritte Stufe handelt mit dem Wert von einzelnen Unternehmungen, die diese Güter erzeugen, auf einer übergeordneten Art von Markt, nämlich der Börse“ (Schwarz, 2012).

Aktien können, müssen aber nicht an einer Börse gehandelt werden. Man unterscheidet hierbei börsennotierte und nicht börsennotierte Unternehmen. Grundsätzlich ist eine Aktie ein Anteil am Unternehmenswert einer Aktiengesellschaft. Der Unternehmenswert berechnet sich aus den erwarteten diskontierten Zahlungsüberschüssen. Diese Zahlungsüberschüsse hängen vom künftigen Erfolg des Unternehmens ab. Um den aktuellen und zukünftigen Unternehmenserfolg zu bestimmen, ist es notwendig alle relevanten Informationen, die Einfluss auf den Unternehmenswert haben könnten, zu kennen. Diese vollkommene Informationslage ist in der Praxis jedoch nie zu erreichen, wird allerdings in manchen Modellen zu Grunde gelegt, um das Verhalten von Börsenmaklern zu erklären. Dieses Konzept wird auch als Informationseffizienz bezeichnet. Wenn man von diesem Konzept ausgeht, bedeutet es, dass alle Marktteilnehmer über alle Informationen verfügen und ihre Entscheidungen danach richten. Der Aktienpreis reflektiert somit zu jeder Zeit alle vergangenen und zukünftigen Informationen vgl. (Fama, 1970). Dieser Umstand erklärt sich folgendermaßen. Um einen Gewinn zu erzielen, versucht ein Akteur die zeitliche Preisdifferenz einer Aktie auszunutzen. Dieses Vorgehen wird als Spekulation bezeichnet. Da

Kaufentscheidungen immer in der Gegenwart, und Verkaufsentscheidungen immer in der Zukunft getroffen werden, wird der Kauf einer Aktie aufgrund einer Annahme über die Zukunft getroffen. Dieses Verhalten führt dazu, dass der Wert einer Aktie nicht den gegenwärtigen Wert des Unternehmens wiederspiegelt, sondern den zukünftigen Wert. Da dieser Wert von den künftigen Gewinnen und Verlusten eines Unternehmens abhängt, spiegelt der aktuelle Aktienkurs die Annahme über künftige Gewinne wieder. Aufgrund dessen, dass Annahmen über die Zukunft immer nur aufgrund aktueller Informationen getroffen werden können, sind Informationen die Haupteinflussgröße für die Preisbildung auf dem Aktienmarkt.

2.1.2 Einfluss der gesellschaftlichen Stimmung auf den Aktienmarkt

Während naturwissenschaftliche Phänomene den Vorteil haben, eine begrenzte Anzahl an Einflussfaktoren zu besitzen und sich daher gut in einem Modell abbilden lassen, sind Aktienkurse weitaus schwieriger zu modellieren. Viele der existierenden Finanzmarkt-Modelle beschreiben meist nur isolierte Teilaspekte des Marktes und schließen das soziale System der Gesellschaft, deren Teil sie sind, aus. Ökonomische Systeme sind jedoch auch immer Teil sozialer Systeme. „The economy is not a physical system. It is a system of human interactions" (Nofsinger, 2005). Während bestimmte Einflussfaktoren den Markt bis zu einem gewissen Grad beschreiben können, ist es vor allem das, was Investoren über diese Faktoren denken, was ihr Handeln bestimmt. Was Investoren über Aktienkurse denken, wird dabei sehr stark durch die Absprache mit anderen Investoren geprägt. Sie treffen ihre Entscheidungen daher nicht alleine, sondern in Interaktion mit anderen in einem sozialen Kontext. „The economy is the sum of the economic interactions in society" (Nofsinger, 2005). Dabei können Emotionen ihre Einstellung sehr stark beeinflussen, auch wenn diese Emotionen nicht direkt mit ihrer aktuellen Entscheidungsfindung zusammenhängen. Beispielsweise führt ein höherer Grad an Optimismus in der Gesellschaft auch zu optimistischeren Investoren. Dieser Optimismus beeinflusst dabei ihre Entscheidungsfähigkeit gegenüber Risiken und Unsicherheit. Ein Überoptimismus führt beim Investor oft zu einer Überschätzung des Erfolgs und zu einer Unterschätzung der Risiken. Bei einer extremen Übereuphorie kann es beispielsweise zu einer Blase im Aktienmarkt kommen. Bei einer pessimistischen Grundeinstellung der Investoren werden diese dagegen eher risikoaversiv handeln, was zu weniger Aktienkäufen und einen sinkenden Kurs führt.

Die Ausbreitung von Informationen geschieht in einem sozialen System meist exponentiell. Jede Person kann ihr Wissen an eine Vielzahl anderer Personen weitergeben und diese wiederum an andere. Informationen können sich hierdurch extrem schnell verbreiten. Nach (McGuire, 1985) ist dabei eine Konversation effektiver als die mono direktionale Verbreitung von Informationen über Massenmedien. Durch den interaktiven Teil einer Unterhaltung werden eher Emotionen stimuliert. Damit eine Information bei einem Menschen eine Handlung auslöst, muss sie des Weiteren überzeugend sein. Informationen werden dabei durch die eigene Beurteilung interpretiert. Menschen müssen Informationen als relevant ansehen, damit diese überzeugend sein können vgl. (McCloskey & Arjo, 1995). Unterhaltungen über den Aktienmarkt und mögliche Investitionen finden dabei durch eine Vielzahl an Personen statt. Broker unterhalten sich mit ihren Klienten und anderen Brokern. Analysten unterhalten sich mit Führungspersonen und Managern. Der einzelne Investor

spricht mit seiner Familie, mit Nachbarn und Freunden. Die Meinung und Einstellung dieser Personen hat daher großen Einfluss auf die finanziellen Entscheidungen des Aktienhändlers.

Bei Entscheidungen versuchen Menschen die Tragweite und die Wahrscheinlichkeit eines Ereignisses vorherzusehen. Dabei schätzt eine Person nicht nur das Ergebnis voraus, sondern auch die eigene Gefühlslage bei Eintreten eines Ereignisses. Dieser Prozess beeinflusst sowohl den kognitiven Evaluationsprozess als auch den aktuellen emotionalen Zustand des Entscheidungsträgers. Es kann vorkommen, dass der emotionale Zustand einer Person den Entscheidungsprozess dominiert und rationale Denkweisen in den Hintergrund treten.

Joseph Forgas hat ein Modell entwickelt, das feststellen sollte, bis zu welchem Grad sich Menschen bei ihrer Entscheidungsfindung auf Emotionen verlassen vgl. (Forgas, 1995). Er stellte dabei fest, dass bei Dingen wie Risiko und Unsicherheit, Emotionen eine wichtige Rolle spielen. Je höher der Grad an Komplexität und Unsicherheit, desto mehr spielen Emotionen eine Rolle im Entscheidungsprozess. John Nofsinger argumentiert daher, dass die soziale Grundeinstellung sich ebenso auf ökonomische Entscheidungen auswirkt vgl. (Nofsinger, 2005). Edward Sanders hat des Weiteren herausgefunden, dass ein Zusammenhang zwischen dem Wetter und den Kapitaleinnahmen der New Yorker Börse besteht vgl. (Sanders, 1993). Ein sonniges Wetter verbessert die Laune der Broker und macht sie optimistischer. An Sonnigen Tagen gibt es seiner Aussage zufolge daher höhere Gewinne. Ein anderes Modell von (Shiller, 1984) zeigte, dass sich eine Gemütslage ähnlich einer Krankheit ausbreitet. Die Rate, mit der Personen mit einer veränderten Gemütslage in Kontakt kommen, ähnelt einer Infektionsrate. Ein neu gefundener Optimismus kann sich daher unter Brokern relativ schnell ausbreiten und sich somit schnell auf den Aktienkurs auswirken.

Die letztendlichen Kauf und Verkaufsentscheidungen der Broker hängen dabei von deren Erwartungshaltung ab. Um die Erwartungshaltung rational fassen zu können, verwenden viele Broker spezielle Analysewerkzeuge. Die Nutzung dieser Werkzeuge erfordert jedoch ebenfalls das Schätzen zukünftiger Ereignisse. Diese menschlichen Einschätzungen sind nun wiederum sehr stark emotionsabhängig. Robert Shiller argumentiert weiterhin, dass die meisten Investoren ohnehin kein Verständnis von Datenanalysen und Risikoanpassungen besitzen und sie kein Modell des Aktienmarktes haben nach dem sie vorgehen vgl. (Shiller, 1984). Wenn die Einstellung und die Entscheidungen der Aktienhändler nun emotionsabhängig sind dann besteht die Möglichkeit, dass der Aktienmarkt als Ganzes durch Emotionen beeinflusst werden kann.

Emotionen verursachen bei Investoren dabei sehr oft folgendes Vorgehen. Ein Investor, der eine im Wert gestiegene Aktie verkauft, empfindet ein Glücksgefühl, da er eine gute Entscheidung getroffen hat. Wenn der Aktienkurs jedoch nach einem Kauf der Aktie fällt, neigt der Investor oft dazu, die Aktie zu halten, in der Hoffnung, dass sie irgendwann wieder steigt. Hiermit will er das schlechte Gefühl umgehen, eine schlechte Entscheidung getroffen zu haben. Diese emotionalen Zustände führen also dazu, dass im Wert steigende Aktien verkauft und im Wert fallende Aktien gehalten werden.

Neben dem regulären Handelsbetrieb auf den Aktienmärkten, gibt es hin und wieder auch Zeiten von öffentlicher Euphorie an den Märkten, auf die oft ein Einbruch der Kurse folgt.

Diese auch als Aktienblasen bekannten Ereignisse, werden oft vom kollektiven Verhalten der Gesellschaft verursacht. Durch diese öffentlichen Manien und Paniken wird das individuelle Urteilsvermögen der Investoren stark beeinflusst. „…they consider manias and panics to be periods in which collective behavior and social mood subverts established individual guides of behavior" (Turner and Killian (1987)). Die Befragung einzelner Investoren ergab, dass diese die Blase, in der sie sich befanden, durchaus erkannten hatten, jedoch hofften sie darauf, dass die Blase noch weiter anhalten würde, und sie somit weiterhin große Gewinne machen könnten. Dieses Verhalten ist ein weiteres Beispiel wie extremer Optimismus der Investoren den Aktienkurs beeinflussen kann. Dabei kann es passieren, dass Investoren sich gegenseitig in ihrer Meinung bestätigen und sich ihre Grundhaltung dadurch weiter verstärkt.

Wenn diese emotionsgeladene euphorische Stimmung an den Märkten nachlässt und die Investoren anfangen ihr Handeln rational zu überdenken, fangen die Kurse wieder an zu sinken, da die Investoren merken, dass die Preise vollkommen überbewertet sind. Die Blase platzt. Wenn rationale Investoren davon ausgehen, dass die Preise des Marktes sich von deren eigentlichen Marktwert entfernen, werden manche versuchen davon zu profitieren. Durch diese Haltung handeln selbst rationale Investoren wie emotionale Investoren. Es zeigt sich also, dass durchaus eine Verbindung zwischen der emotionalen Haltung der Gesellschaft und dem Aktienkurs eines Unternehmens bestehen kann.

2.2 Analysemethoden

Um die im späteren Verlauf dieser Arbeit genutzten Analysemethoden besser verstehen und in ihrem wissenschaftlichen Kontext stellen zu können, werden im Folgenden die Methoden beschrieben, auf denen alle späteren Techniken basieren. Dies sind die Techniken des Data Mining, Web Mining, Text Mining, Opinion Mining und Reality Mining. All diese Techniken beschäftigen sich mit der Analyse von Daten, nutzen allerdings verschiedene Quellen oder verfolgen unterschiedliche Ziele. Daher erfolgt hier eine getrennte Betrachtung. Die Hoffnung aller Analysemethoden liegt im Erkenntnisgewinn aus der Flut von Daten. "Computers have promised us a fountain of wisdom but delivered a flood of data" (Frawley, et al., 1992). Grundlage für derartige Analysemethoden ist die immer weiter steigende Datenmenge sowohl im privaten als auch im geschäftlichen Umfeld. Dies können private Daten wie Telefongespräche, Kreditkarteninformationen oder besuchte Webseiten im Netz sein. Im geschäftlichen Bereich fallen Daten wie Einkauf und Verkaufs-Transaktionen oder Serverlogs an. Diese riesige Datenmenge beinhaltet wertvolle Informationen, die für Unternehmen immer mehr erfolgsentscheidend werden. Die schiere Menge der Daten macht dabei eine manuelle Analyse durch einen Mensch praktisch unmöglich. Derartige Datenmengen können nur noch maschinell mit der Hilfe von mathematisch-statistischer Algorithmen bearbeitet werden. Vor allem bei Banken, Versicherungen, im Handel und der Telekommunikationsbrnache herrscht für derartige Verfahren ein großer Bedarf.

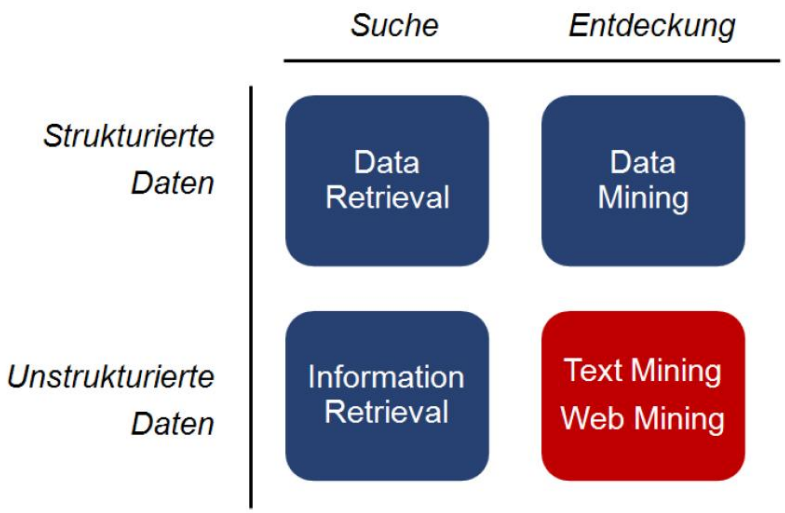

Abbildung 1: Thematische Einordnung der Analysemethoden

Die verschiedenen im Folgenden vorgestellten Methoden unterscheiden sich nun darin, ob die Datenmenge die man untersucht, vollständig bekannt ist, und ob diese in strukturierter Form vorliegt. Abbildung 1 zeigt die verschiedenen Klassen von Analysemethoden nach (Heyer, 2010) . Methoden bei denen die gewünschte Informationsmenge noch nicht vorliegt und man daher auf deren Suche angewiesen ist, werden als Retrieval bezeichnet. Je nachdem ob die gesuchten Daten in strukturierter oder unstrukturierter Form vorliegen, spricht man von Daten oder Information-Retrieval. Das zweite Themengebiet stellt das des Mining dar. Von Mining spricht man, wenn die zu analysierende Datenmenge bereits vorliegt und man versucht daraus neue Erkenntnisse zu gewinnen. Bei strukturierten Daten spricht man von Data Mining, bei unstrukturierten Daten von Textmining. Unstrukturierte Daten können beispielsweise in Form von Fließtexten oder Webseiten vorliegen. Eine Aufbaustufe des Text Mining stellt das Opinion Mining dar. Hier wird versucht durch die Methoden des Text Mining die Meinung einer Aussage zu erkennen. Die hier vorgestellten Analysemethoden werden nun im Folgenden näher beschrieben.

2.2.1 Information Retrieval

Das Information Retrieval (IR) stellt die Basis jeder elektronischen Informationssuche dar. Norbert Fuhr beschreibt IR zusammengefasst als „inhaltliche Suche in Texten“ vgl. (Fuhr, 2010). Diese Definition ist jedoch zu eng gefasst, da sich das IR mit der Suche jeglicher Information beschäftigt, auch wenn diese nicht an Texte gebunden ist. Die Fachgruppe Information Retrieval der Gesellschaft für Informatik fasst den Begriff daher etwas weiter, nämlich als „Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden“ (Gesellschaft für Informatik, 2009). Einfacher ausgedrückt, das Finden von relevanten Informationen. Als Informationsträger können verschiedenste Medien zum Einsatz kommen: Bild, Text, Film usw. Da der größte Teil der weltweit verfügbaren Informationsmenge in Textform vorliegt, beschäftigt sich das IR allerdings fast immer mit der Verarbeitung von Texten.

Seit der Ausbreitung des World Wide Web hat das Information Retrieval enorm an Einfluss gewonnen. Firmen wie Google, deren Suchdienst nichts anderes ist, als eine IR Anwendung, verdeutlichen die heutige Bedeutung dieser Forschungsrichtung. Wie findet jemand jedoch für

ihn relevante Informationen? Das IR sieht hierfür verschiedene Methoden vor. Die Expertenbefragung, bei der, der Informationssuchende sich an andere Personen richtet, die jene Information bereits besitzt, die er sucht. Die Literaturrecherche, bei der davon ausgegangen wird, dass sich die gesuchte Information in einer Sammlung bekannter Literatur befindet. Die Bibliotheksrecherche, bei der in einem Katalog nach relevanter Literatur gesucht wird. Und schließlich die Online Suche, die sich von der Bibliotheksrecherche dahin gehend unterscheidet, dass die Menge und Qualität der verfügbaren Literatur je nach Quelle sehr schwanken kann. Während es vor wenigen Jahrzehnten noch sehr mühsam war überhaupt an eine größere Menge an Informationen zu gelangen, ist spätestens seit Einführung des Webs der Zugang zu Informationen mühelos möglich. Die zu Verfügung stehende Menge an Informationen übersteigt dabei die Menge der für den einzelnen relevanten Informationen bei weitem. Dem Begriff der Relevanz kommt daher eine besonders wichtige Rolle zu. Die Relevanz gibt an „zu welchem Grad ein Dokumente zur Informationssuche passt und zur Befriedigung des Informationsbedürfnisses eines Nutzers beiträgt“ (Gottron, 2010). Nach (Fuhr, 2010) gibt es vier verschiedene Arten von Relevanz. Die situative Relevanz, die ausdrückt, ob ein gefundenes Dokument in einer konkreten Situation als nützlich angesehen wird. Die subjektive Relevanz, die den Nutzen einer konkreten Person zu einem Dokument angibt. Die objektive Relevanz, die versucht durch das Urteil mehrerer unabhängiger Nutzer eine Objektive Relevanz zu einem vorgegebenen Informationsbedürfnis festzulegen. Und schließlich die Systemrelevanz, bei der die Nützlichkeit eines Dokumentes anhand eines maschinellen Algorithmus bestimmt wird.

Das größte Hindernis beim Finden relevanter Dokumente stellt die Vagheit der menschlichen Sprache dar. Diese ist nicht immer exakt, sehr oft kontext- sowie interpretationsabhängig. Vagheit kann sich beispielsweise in Form von Synonymen, Homonymen oder Ironie ausdrücken. Wie definiert man jedoch, ob ein Dokument relevant ist oder nicht? Eine der Grundideen des IR stellt das so genannte Cranfield Paradigma dar. Die Kernidee besteht darin, dass ein fester Korpus aus Dokumenten und ein gegebenes Informationsbedürfnis bestehen. Eine professionelle Jury entscheidet dann darüber, welche Dokumente des Korpus zu diesem Informationsbedürfnis relevant sind oder nicht. Durch die große Anzahl heutiger Informationsquellen wird dieser Vorgang inzwischen meist maschinell durchgeführt. Einer der wichtigsten Begriffe des Information Retrieval ist der des Goldstandard. Dieser definiert die ideale Menge an Dokumenten zu einem bestimmten Suchbegriff. Falls der Goldstandard bekannt ist, kann die Klassifikationsgüte eines IR Systems bestimmt werden, da ein Abgleich mit dem Goldstandard angibt, ob ein Dokument relevant ist oder nicht. Zwei Metriken, die zur Messung der Klassifikationsgüte genutzt werden sind Precision und Recall. Die Precision gibt an, welchen Anteil der Dokumente die eine Suchanfrage zurück gibt, im Goldstandard enthalten und somit relevant sind. Der Recall beschreibt wie groß der Anteil der zurückgelieferten Dokumente im Verhältnis zu allen relevanten Dokumenten ist. Anders ausgedrückt welchen Anteil aller Relevanten Informationen habe ich bekommen.

2.2.2 Data Mining

Das Data Mining beschäftigt sich mit der Analyse von strukturierten Daten, meist in Form von Tabellen. Frawley, Piatetsky-Shapiro und Matheus definieren Data Mining als „die Extraktion und Entdeckung von implizitem, bisher nicht bekanntem und potenziell nützlichem Wissen aus Daten“ (Frawley, et al., 1992). Oft wird Data Mining auch als Knowledge Discovery from Database (KDD) bezeichnet. Diese Begriffe werden meist synonym verwendet. „Daten sind das Ergebnis der Zuordnung von Zeichen zu Ausschnitten der Realität oder Vorstellungswert des Menschen“ (Lehner & Maier, 1994). Daten schaffen den Bezug von der Syntax zur Semantik. Für die Anwendung von Data Mining Verfahren ist dabei entscheidend, dass die Semantik der Daten bekannt ist. Ein Wert ohne seine Bedeutung ist nutzlos. „Entscheidungsunterstützend wirkt Data Mining durch die Generierung von Modellen zur Abbildung von Input-Output-Relationen“ (Peterson, 2005).

Data Mining Analyse Software wie SPSS, SAS Enterprise Miner oder Rapidminer erlauben es auch Personen die kaum Hintergrundwissen in entsprechenden Algorithmen besitzen, Data Mining betreiben zu können, in dem sie vorgefertigte Analysebausteine bereitstellen, die der Anwender als Black Box verwenden kann. Die Benutzeroberflächen dieser Anwendungen werden dabei immer benutzerfreundlicher, was auch unerfahrenen Anwendern den Einstieg erleichtert. Aber welchen Beitrag können Informationen für den Unternehmenserfolg leisten? Einer der wichtigsten Beiträge ist die Hilfe zur Entscheidungsfindung. Entscheidungen werden grundsätzlich aufgrund von Informationen getätigt. Diese betreffen Ziele, Umweltzustände, Alternativen und Konsequenzen. Dabei unterscheidet man zwischen einer vollkommenen Informationslage, bei der alle Informationen die für die Entscheidung Relevanz haben, bekannt und sicher sind. Und einer unvollkommenen Informationslage bei der entweder Informationen fehlen oder ihre Richtigkeit unsicher ist. Dabei ist es für ein Unternehmen überlebensnotwendig über eine Informationslage zu verfügen, die der Konkurrenz mindestens ebenbürtig ist.

Da die Menge der Informationen, die für eine hinreichende Inforationslage notwendig ist, so groß ist, dass sie mit manuellen Methoden nicht zu handhaben ist, werden die maschinellen Analyse Methoden des Data Mining immer wichtiger. Die Anwendungsmöglichkeiten hängen dabei sehr stark von der Branche und der Datenlage ab. Typische Anwendungsbereiche sind beispielsweise die Einteilung von Kunden in Gruppen durch die Analyse von umsatz- und soziodemokrafischen Daten. Dies soll zu einer persönlicheren Ansprache der Kunden und für gezieltere Marketing-Maßnahmen vgl. (Peterson, 2005) führen. Weitere Anwendungen sind die Bonitätsprüfung vor der Kreditvergabe oder die Analyse der Verbindungsdaten bei Telekommunikationsanbietern, für die Erstellung neuer Tarife, die Analyse von Log-Dateien um Hackerangriffe zu erkennen, die Warenkorbanalyse in Off- und Online Shops um Produkte, die oft zusammen gekauft werden, zusammen zu platzieren oder sie als Paket gemeinsam zu verkaufen. Typische Techniken sind hierbei die Segmentierung, die Link-Analyse oder die Klassifikation. Die Segmentierung oder auch Clustering untersucht die Ähnlichkeiten von Dingen anhand von Daten die deren Eigenschaften beschreiben. Es wird versucht, in einer heterogenen Menge von Dingen, homogene Subgruppen zu finden. Die Link Analyse versucht Zusammenhänge zwischen dem Auftreten mehrerer Ereignisse herzustellen um zu untersuchen, ob eine kausale Beziehung zwischen ihnen besteht. Es gilt

also heraus zu finden, ob auf ein Ereignis A, ein Ereignis B folgt und wenn ja mit welcher Wahrscheinlichkeit. Die Klassifikation versucht aufgrund einer Reihe von Input-Variablen eine Zielvariable beziehungsweise eine Zieleigenschaft möglichst gut zu approximieren.

2.2.3 Web Mining

„Web Mining bezeichnet die allgemeine Anwendung von Verfahren des Data Mining auf Datenstrukturen des Internet“ (Zaïane, 2001). Als Datenmaterial dienen hier sehr oft Formate wie HTML oder XML. Diese Formate werden als semistrukturiert angesehen, da sie eine gewisse wenn auch nicht einheitliche Struktur besitzen. Sowohl HTML als auch XML haben zwar eine gewisse Struktur, diese unterscheidet sich allerdings von Quelle zu Quelle. Wie bei den anderen Mining Verfahren dient auch das Web Mining dem Erkenntnisgewinn. „Web Mining is the extraction of interesting and potential useful patterns and implicit information from resources or activity related to the World Wide Web“ (Zaïane, 2001). Man unterscheidet beim Web Mining drei verschiedene Formen.

Web Content Mining

„Web Content Mining beinhaltet die Analyse des Inhalts von Webseiten mit dem Ziel, die Suche nach Informationen im Web zu erleichtern. Aufgaben sind z.B. die Klassifizierung und Gruppierung von Online-Dokumenten oder das Auffinden von Dokumenten nach bestimmten Suchbegriffen“ (Blensberg, 2001).

Web Structure Mining

„Web Structure Mining untersucht die Anordnung von Objekten innerhalb einer Website sowie verschiedene Seiten zueinander. Besonderes Augenmerk liegt hierbei auf den Verlinkungen zwischen den Seiten“ (Srivastava, et al., 2000).

Web Usage Mining

„Web Usage Mining analysiert das Verhalten von Besuchern auf Websites. Hier werden mit Hilfe von Data Mining-Methoden Logfiles ausgewertet, Verhaltensmuster von Internet Nutzern identifiziert und einer Klassifikation unterzogen. Sehr aufschlussreiche Analyseergebnisse können auch über Assoziationen erreicht werden, die analog zu Warenkorbanalysen hier Sequenzen betrachten“ (Peterson, 2005).

Die spätere Analyse von Twitter-Nachrichten zur Vorhersage von Börsenkursen kann dem Web Content Mining zugeordnet werden, da es sich mit dem von Usern erstellten Inhalt einer Website auseinandersetzen wird. Als Datenquellen fungieren beim Web Mining vor allem Web Crawler. Dies sind Programme die das Internet nach Informationen durchsuchen. Dabei durchsuchen sie Webseiten nach Hyperlinks und nutzen die dort referenzierten Seiten als neuen Ausgangspunkt, um die Ergebnismenge auszuweiten. Es kann jedoch nicht sichergestellt werden, dass ein Crawler alle relevanten Informationen findet. Kein Programm wird jemals ein vollkommen perfektes Ergebnis liefern können. Desweiteren haben Web Crawler das Problem, dass sie womöglich nie zu einem Ende kommen, da immer neue verlinkte Webseiten ins Netz gestellt werden.

Der Ablauf einer Web-Mining-Analyse kann prinzipiell in drei Schritte zerlegt werden. Erstens dem finden von Webseiten, die relevante Informationen enthalten. Zweitens dem extrahieren von Informationen aus diesen Seiten, die für die eigene Untersuchung von Interesse sind. Und schließlich dem dritten Schritt, bei dem die extrahierten Informationen dann zueinander in Verbindung gesetzt werden, um daraus neue Erkenntnisse herzuleiten.

2.2.4 Text Mining

Im Gegensatz zum Data Mining beschäftigt sich das Text Mining mit der Analyse unstrukturierter Texte. Nach Marti Hearst versteht man darunter „...alle Techniken zum Entdecken und automatischen Extrahieren von neuen, zuvor unbekannten Informationen aus Texten“ (Hearst, 2003). Es findet nach dieser Definition allerdings keine Eingrenzung statt, um welche Art Text es sich handeln kann. Jede Form von Text sowohl elektronisch als nicht elektronisch fällt somit als Quelle in diese Kategorie. Nicht elektronische Texte wie Papier Dokumente können dabei in eine elektronische Form überführt werden. Ein grundlegendes Problem bei der automatisierten Analyse von Text liegt in seiner Unstrukturiertheit. Um derartige Daten verarbeiten zu können, ist eine Aufbereitung des Datenmaterials notwendig. Ziel ist es die Daten in ein klar strukturiertes Schema zu überführen, das über alle Datensätze hinweg gleich bleibt. Für die Vorgehensweise einer Text-Mining-Analyse sehen (Hajo & Rentzmann, 2006) folgende Schritte vor.

1. Aufgabendefinition

2. Dokumentselektion

3. Dokumentaufbereitung

4. Text-Mining-Methoden

5. Interpretation und Evaluation der Ergebnisse

6. Anwendung der Ergebnisse

Der erste Schritt besteht darin zu definieren, was man eigentlich herausfinden möchte. Es wird also ein Ziel gesetzt, dessen Erreichen messbar ist. Im zweiten Schritt wird aufgrund der Zielvorgabe eine Menge an relevanten Dokumenten ermittelt. Nachdem eine Menge von Dokumenten vorliegt, müssen diese in Schritt drei in eine strukturierte Form überführt werden. Hierzu werden verschiedene Merkmale aus dem Text extrahiert und alle Dokumente in eine einheitliche Form gebracht. Die hier am meisten verwendeten Methoden sind die Tokenisierung, das Part of Speach Tagging und die Lemmatisierung. Die Tokenisierung versucht Wörter aus einem Fließtext zu erkennen. „Unter Part-of-speech Tagging versteht man die Zuordnung von Wörtern und Satzzeichen eines Textes zu Wortarten (engl. part of speech). Hierzu wird sowohl die Definition des Wortes als auch der Kontext (z.B. angrenzende Adjektive oder Nomen) berücksichtigt“ (De Choudhury, et al., 2010). Die Lemmatisierung versucht alle Wörter in ihre grammatikalische Grundform umzuwandeln.

Erst nach dieser Vorverarbeitung kann in Schritt vier die eigentliche Analyse stattfinden. Die hier angewendeten Methoden entsprechen größtenteils denen des Data Mining. In Schritt fünf werden die durch die Analyse gewonnen Ergebnisse interpretiert und auf ihre Gültigkeit hin

evaluiert. Wenn die hier gewonnenen Erkenntnisse der Zielsetzung noch nicht genügen, wird ein Rücksprung auf Schritt vier getätigt und die Text Mining Methoden entsprechend angepasst. Wenn die Ergebnisse der Analyse der Aufgabenstellung gerecht werden, folgt die Anwendung der Erkenntnisse. Mit diesem Schritt terminiert der Text Mining Prozess.

2.2.5 Opinion Mining

Im Gegensatz zu den faktenbasierten Analysemethoden des Data Mining und Web Mining, beschäftigt sich das Opinion Mining mit der Untersuchung subjektiver, personenbezogener Meinungen. Diese Art der Analyse findet seit der enormen Verbreitung persönlicher Äußerungen im Internet durch die verschiedenen Plattformen des sogenannten Web2.0[2] immer mehr Beachtung. Vor allem für Unternehmen stellt die riesige Fülle an Meinungsäußerungen eine große Chance dar, die Einstellung der Kunden zu ihren Produkten und die der Konkurrenz zu erfahren. Während in der Vergangenheit hierfür persönliche Befragungen notwendig waren, beispielsweise mit Hilfe von Fragebögen oder Telefongesprächen, sind derartige Daten heute bereits in Netz vorhanden. Viele Firmen haben dabei erkannt, dass die Meinungen der Konsumenten enormen Einfluss auf andere Menschen haben und somit ihre Kaufentscheidung und Markentreue beeinflussen vgl. (Liu, 2009). Die Vielzahl an Einzelmeinungen macht eine vollständige Analyse aller Meinungen durch eine manuelle Bearbeitung dabei schlichtweg unmöglich. Obwohl ein Mensch bei der Bewertung von Meinungen einer Maschine bei weitem überlegen ist, macht es daher Sinn für derartige Untersuchungen einen maschinellen Ansatz zu wählen. Obgleich die Qualität der Einzelbewertung hierdurch sinkt, erlaubt die gestiegene Quantität eine repräsentativere Gesamtaussage. Einige Begriffe werden oft Synonym mit Opinion Mining verwendet. Beispielsweise beschreiben die Begriffe Sentiment Analyse, Subjektivitätsanalyse oder Review Mining meist denselben Umstand. Für die Analyse von Texten wird beim Opinion Mining auf ein Modell zurückgegriffen, das alle beinhalteten Elemente klassifiziert. Dieses Modell beinhaltet die Elemente: Objekt, Features, Meinungsäußerung, Meinungseigentümer und die Semantische Orientierung (siehe Abbildung 2). Ein Objekt stellt eine Entität dar, auf die sich die Meinungsäußerungen beziehen. Dies können unter anderem Personen, Gegenstände oder Firmen sein. Dabei können sich die Äußerungen entweder direkt auf die Entitäten beziehen, oder auf deren Eigenschaften. Diese einzelnen Komponenten samt deren Eigenschaften werden als Features bezeichnet. Eine Meinungsäußerung stellt eine subjektive Bewertung einer Person gegenüber einem Objekt und gegebenenfalls deren Features dar. Die Person die diese Bewertung abgibt wird als Meinungseigentümer bezeichnet. Die Ausrichtung der Meinungsäußerung wird semantische Orientierung genannt. Sie gibt an, ob eine Aussage positiv, negativ oder neutral zu bewerten ist.

[2] „Letztendlich ist aber Web 2.0 ein Begriff unter dem sich verschiedene technologische, soziale, aber auch ökonomische Entwicklungen im Internet ansiedeln. Daher ist es so gut wie unmöglich, eine klare und abgegrenzte Definition zu finden. Klar ist, dass diese Entwicklungen das Internet interaktiver bzw. partizipativer gemacht haben, was in einer intensiveren Zusammenarbeit zwischen den Nutzern resultiert“ (Messerschmidt, et al., 2010). Diese Autoren geben folgende Eigenschaften von Web 2.0 Technologien an. Diese können demzufolge folgende Eigenschaften enthalten: netzbasiert, dynamisch, Inhalte werden von Nutzern erstellt, besitzen ein Rechtesystem das die Einsicht und Veränderbarkeit von Inhalten regelt, Nutzer besitzen eigene Dateien deren Inhalt und Layout sie verändern können, Nutzer können Inhalte anderer User kommentieren und es besteht oft ein Community Gedanke.

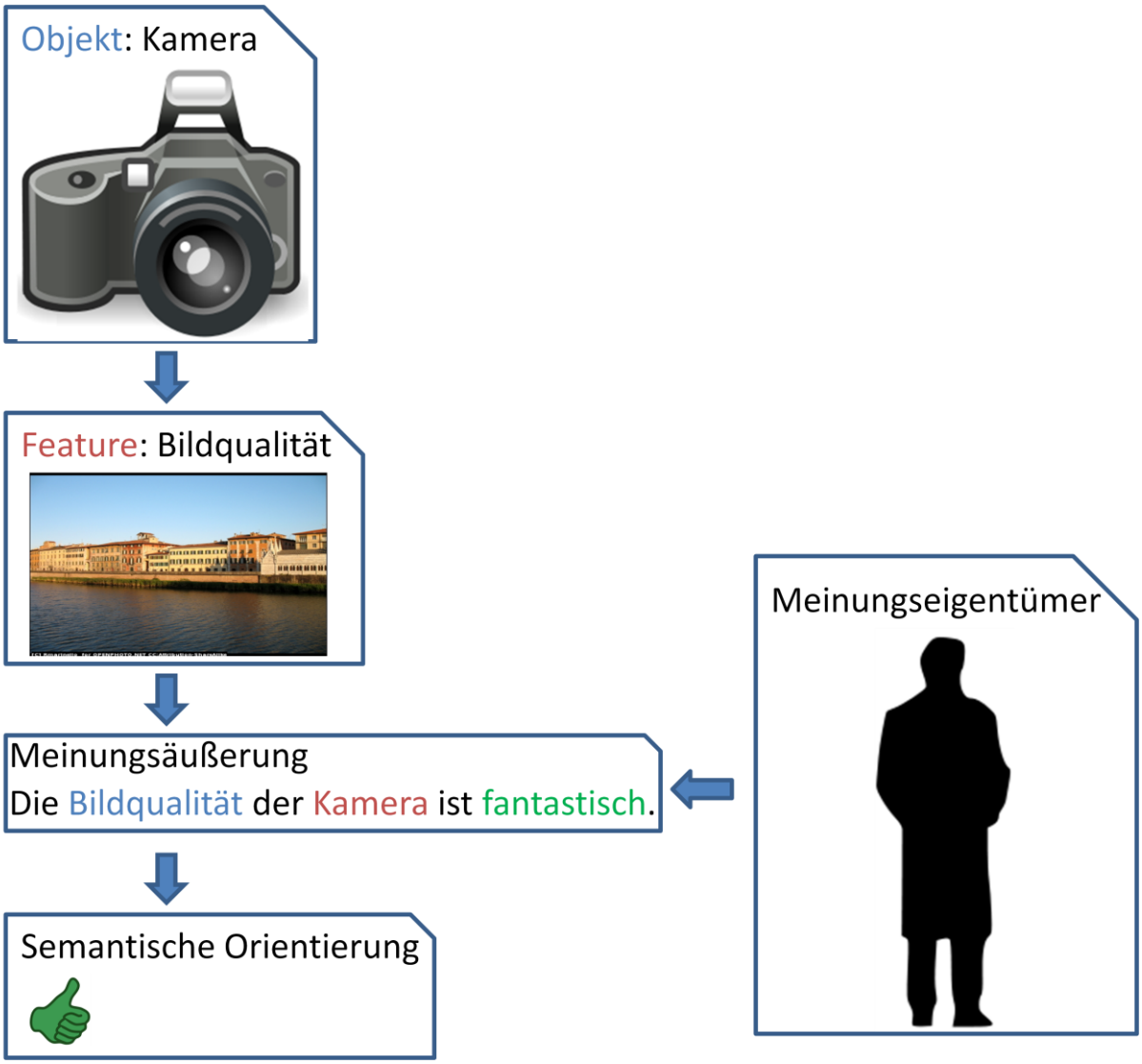

Abbildung 2: Opinion Mining Modell

Abbildung 2 zeigt ein mögliches Beispiel eines Opinion Mining Modells. Das Objekt stellt in diesem Fall eine Kamera dar. Features einer Kamera können unter anderem die Bildqualität, die Größe des Fotospeichers, die Batterielaufzeit oder die Zoomstufen sein. Jedes dieser Features kann nun von einer Person bewertet werden. Bei der Äußerung einer Meinung wird diese Person zum Eigentümer dieser Meinung. Jede Äußerung besitzt dabei ihre eigene semantische Orientierung, die ihn diesem Fall durch das Wort „fantastisch“ positiv ausfällt. Für die Gesamtbewertung eines Objektes können die Meinungen dann entweder pro Feature oder nur für das Objekt selbst aggregiert werden.

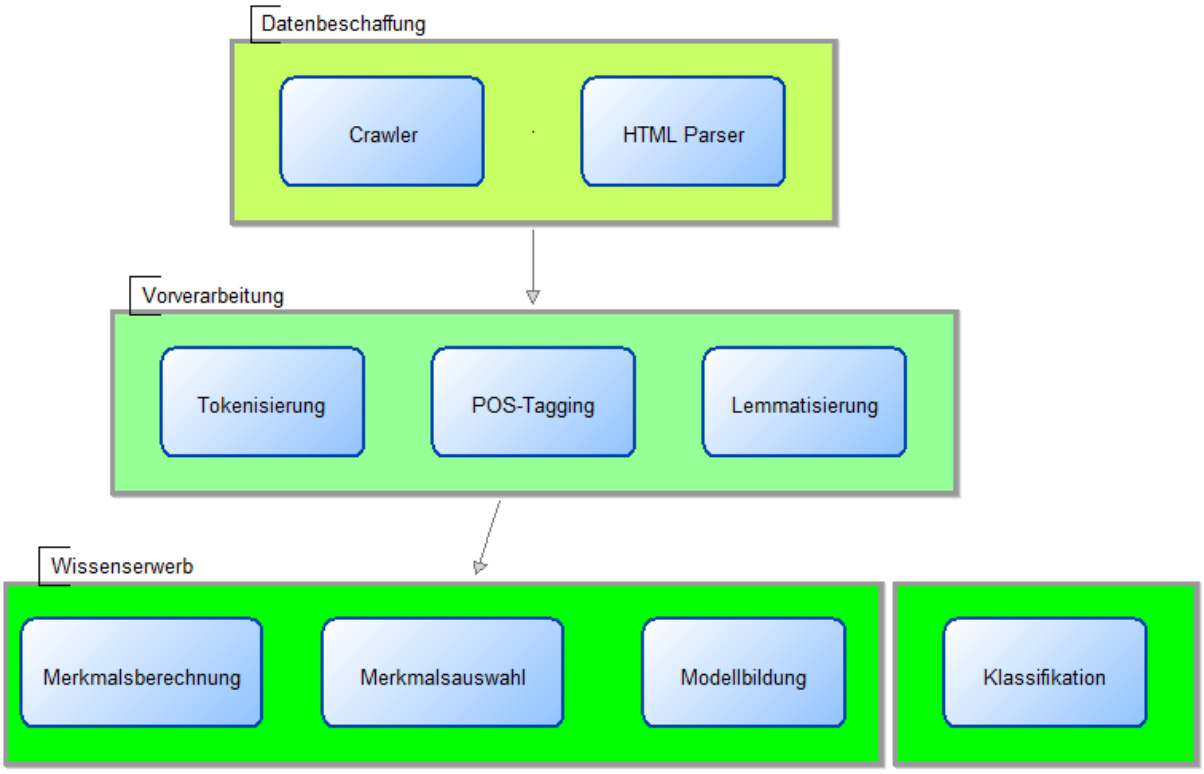

Abbildung 3: Opinion Mining Prozess

Der Ablauf einer Opinion Mining Analyse ähnelt dem des Text Mining, die sich grundsätzlich in Datenbeschaffung, Vorverarbeitung und Wissenserwerb aufteilt (siehe Abbildung 3). Die Datenbeschaffung erfolgt wie beim Web Mining mit Hilfe von Web Crawlern, HTML Parsern oder speziellen APIs, die von einzelnen Webseiten als Schnittstelle zur Verfügung gestellt werden.

Wie beim Textmining werden auch beim Opinion Mining Fließtexte analysiert. Dies bedeutet, dass wiederum unstrukturierte Texte als Datenquelle dienen. Da hierbei allerdings oft netzbasierte Quellen als Datenmaterial dienen, können die Quellen auch Formatierungsinformationen enthalten. Ein klassisches Beispiel hierfür ist das Parsen von HTML Webseiten. Derartige Formatierungsinformationen müssen vor der Analyse entfernt werden, da sie nicht Teil des Inhaltes sind. Am Ende sollten nur noch die Nutzdaten in die Analyse mit eingehen. Die weitere Aufbereitung des daraus resultierenden Fließtextes hängt dann stark von den späteren Analysemethoden ab. Einige Methoden benötigen keinerlei Aufbereitung, andere verwenden Methoden des Text Mining um den Text in eine einheitliche Form zu bringen. Der Teil des Wissenserwerbs stellt beim Opinion Mining die Klassifikation der Meinungsausrichtung dar. Es ist genau dieser Teil der als Sentiment Analyse bezeichnet wird und als Kernstück des Opinion Mining angesehen werden kann.

2.2.6 Reality Mining

Das Reality Mining versucht sozialwissenschaftliche Effekte durch die Analyse personenbezogener Daten zu erklären. Nach Eagle und Pentland definieren es sich als: „ …. the collection of machine-sensed environmental data pertaining to human social behavior" (Eagle & Pentland, 2006). Die immer größere Anzahl an personenbezogenen Daten ermöglich hier einen immer genaueren Einblick in das Privatleben der User. Hierfür werden Daten genutzt die entweder automatisch anfallen, wie beispielsweise Log Daten von Servern oder Daten die von den Usern selbst erstellt wurden, wie Nachrichten von Twitter, Skype oder Facebook. Die vorhandene Datenlage bildet die Realität hierdurch immer mehr ab. Neben der reinen Datenlage sind es jedoch immer ausgereiftere Visualisierungsanwendungen, die es ermöglichen die riesige Datenmenge in eine anschauliche und leicht verständliche Form zu aggregieren. Dienste wie Daytum[3] erlauben dem User sein komplettes Privatleben zu erfassen, es mit anderen Usern zu teilen und es grafisch darzustellen. Obwohl sich diese Anwendung wie die totale Überwachung auswirk, hat der Dienst bereits über 80000 User. Die gesammelten Daten reichen von wissenschaftlichen Dingen bis hin zu banalen Alltagstätigkeiten. Vor allem Marketing- und Werbefirmen interessieren sich für derartige Informationen. Es ist anzunehmen, dass es Firmen geben wird, die Menschen dafür bezahlen werden, gewisse Datenprofile von sich zu erstellen und diese freizugeben. Mit den richtigen Anreizen werden immer mehr Personen dabei gewillt sein, Informationen über sich Preis zu geben. Start-Ups wie Locately[4] bieten ihren Kunden beispielsweise Sonderangebote an, wenn sie sich bereit erklären, ihre Verhaltensdaten weiterzugeben. Jedoch auch ohne Einwilligung der User werden schon jetzt Daten weiter gegeben. Zwar in anonymisierter Form, jedoch lässt sich aus den Datensätzen mit etwas Aufwand eine recht genaue Schätzung aufstellen, zu welcher Person die Daten gehören könnten. Mobilfunkbetreiber geben beispielsweise anonymisierte Verbindungsdaten ihrer Kunden samt Ortsdaten weiter. Wenn man diese Daten betrachtet, kann man den Wohnort einer Person mit sehr hoher Wahrscheinlichkeit vorhersagen in dem man die Ortsdaten des Mobilfunkgerätes in einer Zeit zwischen spät abends und früh morgens betrachtet. Käufer dieser Daten können hiermit umfassende Nutzerprofile erstellen. Auswertungen, die sonst nur der Polizei oder Geheimdiensten zugänglich waren, sind heute in der Hand von privaten Unternehmen. Neben Werbefirmen interessieren sich auch Städteplaner für solche Daten. Sie wollen hierdurch die Verkehrsinfrastruktur besser an die Menschenbewegungen anpassen. Ziel ist es hierbei eine Verkehrsinfrastruktur zu schaffen, die zu weniger Staus und sichereren Großveranstaltungen führt. In Belgien wurde bei einer Untersuchung der Handynutzer herausgefunden, dass die französisch sprechenden und die niederländisch sprechenden Belgier kaum telefonischen Kontakt zu der jeweils anderssprachigen Bevölkerungsgruppe haben. Solche Erkenntnisse dürften sich auf die ohnehin zerrissene Regierung nicht gerade positiv auswirken. Ob und wie solche Dienste umgesetzt werden können, hängt jedoch auch von den Datenschutzbestimmungen des jeweiligen Landes ab. Durch die lasche Datenschutzgesetzgebung auf diesem Gebiet, sind in den USA derartige Dienste am weitesten verbreitet. In Deutschland sind viele dieser Informationen dagegen entweder gar nicht, oder nur unter hohen Auflagen zugänglich.

[3] http://daytum.com/
[4] http://www.locately.com/

3 Mikroblogging & Twitter

Bevor wir Bloggingtexte als Datenquelle nutzen, sollten wir klären was Blogging überhaupt ist, wie dieses sich zusammensetzt, welche Menschen diese Medienform nutzen und zu welchem Zweck. Im Fokus steht hier vor allem die Nutzerschaft von Twitter, da diese das später genutzte Datenmaterial bereitstellt.

3.1 Blogging

„Ein Blog oder auch Web-Log, Wortkreuzung aus engl. World Wide Web und Log für Logbuch, ist ein auf einer Website geführtes und damit meist öffentlich einsehbares Tagebuch oder Journal, in dem mindestens eine Person, der Web-Logger, kurz Blogger, Aufzeichnungen führt, Sachverhalte protokolliert oder Gedanken niederschreibt“ (Wikipedia, 2012). Das Word „log“ (englisch für Klotz oder Holzklotz) stammt in dieser Verwendung aus der nautischen Sprache und wurde früher zur Geschwindigkeitsmessung auf Schiffen verwendet. Man ließ eine Schnur, an der Holzklötze befestigt waren, ins Wasser und zählte die Anzahl an Klötzen die die Strömung in 30 Sekunden wegtrieb. Diese Messung wurde dann in das Logbuch notiert. Das Logbuch diente also der Navigation. Die ersten Blogger verwendeten diese Analogie, da Blogs in dieser Zeit sehr oft als Navigationshilfen für das World Wide Web verwendet wurden vgl. (Rettberg, 2008). Anfang der Neunziger wurde das Word Weblog noch als ein Server Logbuch verstanden, dass die Anzahl an Besuchern einer Webseite sowie deren Verbindungsdaten protokollierte. Einer der ersten, der das Wort Weblog in seinem heutigen Sinn verwendet hat, war Jorn Barger im Jahr 1997. Dieser betrieb eine Webseite, auf der er eine Liste an Links zu Webseiten, die er als nützlich empfand, veröffentliche. Diese Liste wurde von ihm kontinuierlich erneuert. Dieser Seite gab er den Namen „Robot Wisdom a Weblog by Jorn Barner“ vgl. (Rettberg, 2008).

Im Unterschied zu traditionellen publizistischen Methoden, bietet das Blogging dem Autor die Möglichkeit, die vollkommene Kontrolle über den Inhalt und das Layout seines Textes zu behalten. In Zeitungs- und Zeitschriftenartikeln wird dagegen der Text des Autors von einem Editor überarbeitet und im Layout angepasst. Dies schränkt die Meinungsfreiheit des Autors erheblich ein, da er befürchten muss, dass zu kontroverse Äußerungen eine Zensur oder Ablehnung zur Folge haben können. Auf der anderen Seite führt die Redefreiheit des Bloggers jedoch auch dazu, dass Personen alles veröffentlichen können, ganz gleich des Wahrheitsgehaltes ihres Textes. Allerdings steht der Autor auch immer mit seinem Namen für den Inhalt gerade.

Wichtige Unterschiede zu textgebundenen Medien sind die Möglichkeit in einem Blog Links zu anderen Webseiten zu verwenden und Inhalte in Echtzeit veröffentlichen zu können. Ein Blog ist dabei nicht dazu gedacht, eine endliche Geschichte zu erzählen, sondern einen neuen Gedanken oder Erfahrungen des Autors wiederzugeben. Die Motivation einer Person einen Blog zu betreiben kann dabei unterschiedliche Gründe haben. Bonni Nardi u.a. haben in ihrer Studie 5 Hauptgründe hierfür finden können vgl. (Nardi, et al., 2007). Als Tagebuch, um das eigene Leben zu dokumentieren, als Plattform zur Meinungsäußerung und Kommentierung, der Offenlegung der eigenen Gefühle, um eigene Ideen niederzuschreiben oder eine Gemeinschaft aufzubauen beziehungsweise aufrecht zu erhalten.

3.2 Mikro Blogging

Mikro Blogging gilt als eine Unterform des regulären Bloggings. Die Unterschiede liegen in der begrenzten Länge der Nachrichten, die meist auf 140 bis 200 Zeichen begrenzt sind und den sozialen Netzwerk Mechanismen dieser Dienste. User können Nachrichtenkanäle anderer User abonnieren und so deren gesendeten Nachrichten empfangen. Es wird somit eine direkte Verbindung zwischen Autor und Leser hergestellt. Diese Struktur ähnelt sehr stark der des Instant Messaging bei der eine Peer to Peer Verbindung zwischen zwei Personen aufgebaut wird. Beim Mikro Blogging werden allerdings alle Nachrichten über einen zentralen Server geleitet. Es besteht also keine Peer to Peer Architektur. Der Vorteil einer derartigen Architektur liegt in der vollständigen Dokumentation des Datenverkehrs. Viele sehen Mikro Blogging dabei als das Posten von öffentlichen Nachrichten in Echtzeit an. Der Faktor des Öffentlichen spielt hier eine sehr große Rolle. Im Gegensatz zum Instant Messaging kann jeder die Nachrichten einer anderen Person abonnieren und somit empfangen. Hierdurch werden öffentliche Diskussionen ermöglicht.

Der größte Teil des Mikro Bloggings beschränkt sich auf das Versenden von Textnachrichten, obwohl auch einige Dienste für Audio und Video Nachrichten existieren. Der Netzwerk Effekt macht es außerdem zu einer Social-Software[5]. Der Vorteil gegenüber einem regulären Blog ist die Geschwindigkeit mit der Nachrichten gepostet werden. Durch die kurze Länge können Nachrichten sehr schnell verfasst werden. Sehr aktive Mikroblog Nutzer schreiben viele Nachrichten pro Tag anstatt nur einige pro Woche wie in einem regulären Blog. Hinzu kommt die Nutzung mobiler Mikro Blogging Programme durch die starke Verbreitung von Smartphones. Diese ermöglichen das Senden von Nachrichten von jedem beliebigen Punkt.

Durch die beschränke Länge der Nachrichten eignet sich Mikroblogging nicht für ausführliche Erläuterungen. Diese Art der Information wird daher auch als Mikro Content bezeichnet. Koch und Richter definieren Mikro Content als „Informationsschnipsel, die man glaubt irgendwann mal gebrauchen zu können“ (Richter & Koch, 2007). Diese Schnipsel beinhalten meist kurze Gedanken, Meinungen oder verweisen auf andere Quellen. Der Ursprung von Mikro Blogging liegt in der Status-Update-Funktion[6] von Facebook. Diese sollte den Usern die Möglichkeit geben das zu schreiben was sie gerade tun. Diese Funktion wurde von Facebook 2006 eingeführt.

[5] “Social-Software-Anwendungen unterstützen als Teil eines soziotechnischen Systems menschliche Kommunikation, Interaktion und Zusammenarbeit. Dabei nutzen die Akteure die Potenziale und Beiträge eines Netzwerks von Teilnehmern“ (Tochtermann & Back, 2009).

[6] „…in essence it´s a way to describe what you´re doing or thinking about at the moment. It´s a snapshot into your life, posted as a short text on the Facebook site” (Miller, 2011).

3.3 Twitter

Twitter ist derzeit der größte Anbieter für Mikro Blogging der Welt. Der Dienst ermöglicht seinen Usern das Versenden von 140 Zeichen langen Nachrichten und das Abonnieren von Nachrichtenkanälen anderer User. Die Webseite emarketer.com geht davon aus, dass Twitter 2011 Werbeeinnahmen von 139.5 Millionen Dollar erzielt hat. Gegründet wurde Twitter 2006 als Forschungsprojekt von Jack Dorsey, Biz Stone und Evan Williams bei der Firma Odeo. Ein Grundgedanke von Jack Dorsey war dabei die Nutzung mobiler Geräte zur Kommunikation mittels Text. „I want to have a dispatch service that connects us on our phones using text" (Sagolla, 2009). Die Limitierung der Nachrichtenlänge auf 140 Zeichen resultierte damals daraus, dass der Dienst in der Testphase SMS zur Datenübertragung nutzte. Um eine Nachricht nicht auf mehrere Nachrichten verteilen zu müssen, wurde die Länge soweit beschränkt, dass jede Nachricht in eine SMS passte. Die Grundstruktur von Twitter ist die eines sozialen Netzwerkes. Allerdings mit nur unidirektionalen Verbindungen. Jeder Nutzer kann einem anderen Nutzer folgen (Following). Hierdurch bekommt er alle gesendeten Nachrichten dieser Person zugesendet. Lediglich 22% vgl. (Poblete, et al., 2011) der Nutzer folgen sich hierbei gegenseitig und können somit beidseitig kommunizieren. Nachrichten werden von Twitter anti-chronologisch sortiert und erlauben somit eine zeitliche Nachvollziehbarkeit. Twitter Nachrichten werden auch als Status Updates bezeichnet, da Twitter die Frage an seine Nutzer richtet „what are you doing?". Die Nutzer werden also dazu angehalten ihren derzeitigen Status beziehungsweise Situation zu beschreiben.

Der Netzwerk Charakter von Twitter führt zu einem Informationsaustausch zwischen fremden Menschen. Instant Messeging und Email Dienste sind zwar genauso zum Senden von Nachrichten in der Lage, allerdings zählen derartige Dienste für die meisten Personen zur Privatsphäre und sie wollen sicherlich nicht, dass jeder beliebige Mensch all ihre Emails lesen kann. Durch den öffentlichen Charakter von Twitter kann der Dienst die Selbstorganisation und die Nachrichtenverbreitung von Gruppen unterstützen. Diskussionen über bestimmte Themen lassen sich mit Hilfe von Hashtags realisieren. Dies sind Wörter denen ein Raute Symbol (#) vorangestellt wird. Damit lässt sich ein Tweet zu einem Thema klassifizieren. Jeder der dieser Diskussion folgen oder etwas dazu beitragen möchte, braucht lediglich nach dem korrespondierenden Hashtag zu suchen beziehungsweise es in seine Nachricht zu integrieren. Die Weiterleitung von Nachrichten wird mit so genannten „Retweets" realisiert. Hierdurch können Nachrichten anderer Personen in den eigenen Nachrichtenstrom integriert und so exponentiell verbreitet werden.

Twitter stellt eine öffentliche API zu Verfügung, die es erlaubt, Fremdanwendungen an den eigenen Service anzubinden. Die Offenheit hat eine Vielzahl an Programmen für fast alle Plattformen hervorgebracht und so erst zur massiven Verbreitung von Twitter geführt. Josh Cartone entdeckte bei der Untersuchung von Twitter Nachrichten, dass diese von 142 verschiedenen Client-Anwendungen stammten vgl. (Catone, 2008). Die Mehrheit (56%) der User verwendet dabei die Weboberfläche von Twitter. Die nächst größeren Anwendungen sind Instant Messaging (8%), Twhirl (7%), Twitterific (7%) und SMS-Nachrichten (5%).

Inzwischen nutzen nicht nur Privatpersonen Twitter, sondern auch große Nachrichtensender und Politiker. Bestes Beispiel für einen politischen Einsatz war sicherlich der US Wahlkampf von Barack Obama 2008. Nachrichtenseiten verwenden Twitter meist dazu, um Links zu ausführlicheren Artikeln zu verbreiten. Aber auch Firmen haben Twitter für sich entdeckt. Sie nutzen den Dienst meist für Marketingkampagnen oder um aktuelle Angebote zu bewerben.

Abbildung 4 zeigt den Verlauf des Wachstums an Twitter Nachrichten zwischen Januar 2007 und Januar 2010. Man kann erkennen, dass Twitter 2007 noch kaum Beachtung fand. Im folgenden Jahr wuchs die Nachrichtenmenge jedoch bereits allmählich auf eine Zahl von mehreren Millionen pro Tag. Seit Anfang 2009 zeigt Twitter dann ein extremes Wachstum auf. Der Großteil des Gesamtwachstums von Twitter entfällt somit auf die letzten drei Jahre. Jegliche Phänomene die mit der Nutzung von Twitter zusammenhängen stammen also aus dieser Zeit. Die wissenschaftlichen Untersuchungen auf diesem Gebiet sind daher noch ein sehr junges Forschungsfeld. Die Anzahl akademischer Arbeiten zu diesem Thema steigt seit dem enormen Erfolg von Twitter jedoch immer weiter an.

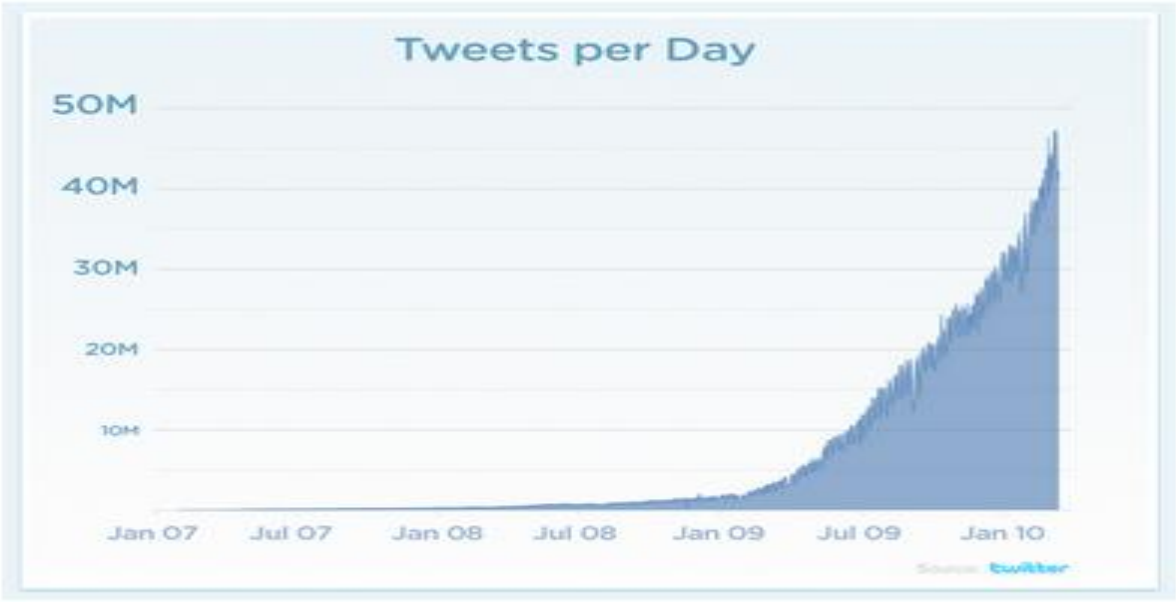

Abbildung 4: Tweets pro Tag 2007 - 2010[7]

[7] Quelle: http://www.marketinggum.com/twitter-statistics-2011-updated-stats/

Dieser Aufwärtstrend setzt sich bis zum heutigen Tag fort und bislang ist noch kein Ende des Booms in Sicht, wie Daten der Webseite compete[8] mit den US Zahlen für den November 2011 gezeigt haben. Deren Werte bescheinigen Twitter ein Wachstum von 45,7% in den letzten 12 Monaten. Nach den Daten dieser Seite rangiert Twitter somit mit über 38 Millionen Seitenzugriffen auf Platz 24 der meist genutzten Websites der USA.

Die Verteilung der in Twitter genutzten Sprachen, zeigt wie zu erwarten, das Englisch mit einem großen Vorsprung die am meisten verwendete Sprache darstellt. Mit Portugiesisch und Japanisch sind die zweit und dritthäufigsten verwendeten Sprachen. Verschiedene Forscher kamen hierbei jedoch zu leicht unterschiedlichen Ergebnissen. Lichan Hong u.a. kamen in ihrer Veröffentlichung von 2011 zu folgendem Ergebnis: Englisch 51,1%, Japanisch 19,1% und Portugiesisch 9,6% vgl. (Hong, et al., 2011). Ein Grund hierfür könnte das unterschiedlich schnelle Wachstum von Twitter in den verschiedenen Ländern sein. Hierdurch wäre die Sprachverteilung einem kontinuierlichen Wandel unterworfen. Bei dem derzeitigen Wachstum von Twitter wird sich diese Entwicklung vermutlich noch einige Zeit fortsetzten. Eine Konsolidierung wird hier sicherlich erst mit einem Ende des Twitter Booms eintreten. Ein weiterer Grund für die unterschiedlichen Ergebnisse könnten die verwendeten Klassifikationsmethoden sein. Da hier eine automatisierte Einteilung vorgenommen wurde, hängen die Ergebnisse vom eingesetzten Algorithmus ab. Die Limitierung der Nachrichtenlänge sowie Unzulänglichkeiten in der Sprachführung führen hier oft zu falschen Ergebnissen. „…common use of slang, along with misspellings, makes automatic language identification particularly challenging" (Poblete, et al., 2011). Vor allem die häufige Verwendung von URLs und Hashtags in den Nachrichten kann die Zuordnung leicht verfälschen. Durch die riesige Anzahl an Twitter-Nachrichten pro Tag ist es desweiteren kaum möglich alle Nachrichten in die Analyse mit einfließen zu lassen. Alle Untersuchungen haben daher lediglich Stichproben verwendet, was das Ergebnis nicht unwesentlich beeinflussen kann.

Wenn man sich die prozentualen Anteile der einzelnen Länder in Abbildung 5 ansieht, korrespondieren diese nur teilweise mit den Ergebnissen der Sprachanteile. Zwar kommen die USA, Großbritannien, Kanada und Australien zusammen auf 43,4%, was zusammen mit den englischen Anteilen anderer Länder vor allem Indien ungefähr 51,1% ergeben könnte. Allerdings passt der hohe Anteil an portugiesischen Tweets hier nicht ins Bild. Portugiesisch wird als Hauptsprache lediglich in Portugal, Brasilien und einigen afrikanischen Ländern gesprochen. Brasilien nimmt jedoch lediglich Platz sechs der Platzierung ein, während Portugal es erst gar nicht unter die ersten zehn Ländern geschafft hat. Die (Group Studie Miniwatts Marketing, 2011) listet desweiteren Portugiesisch mit 3,9% nur auf Platz 5 der Sprachen mit den meisten Internetzugriffen weltweit. Es liegt hier also nahe, dass die Zuordnungen der Tweets zu dieser Sprache fehlerhaft sind.

[8] http://compete.com/us/

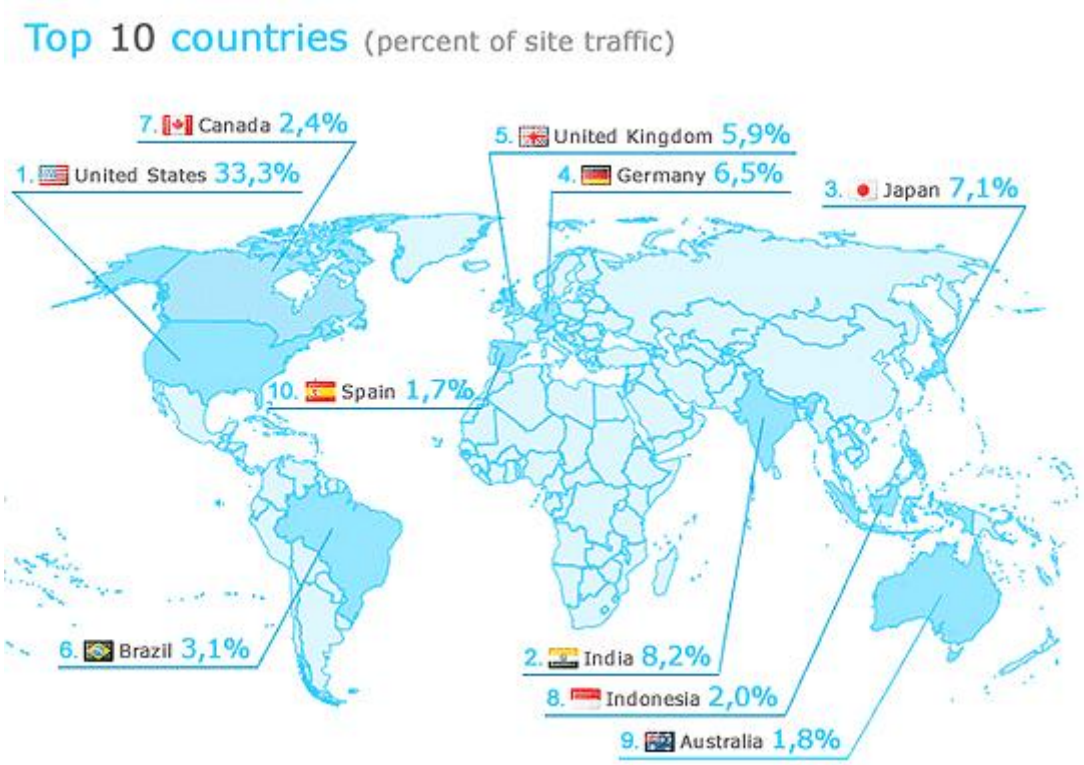

Abbildung 5: Twitter Traffic nach Ländern[9]

Die Miniwatts Studie listet desweiteren Chinesisch mit 24.2 % als die am zweithäufigsten verwendete Sprache im Internet nach Englisch auf. Der Grund warum diese in den Top Ten der Twitter Sprachen nicht aufgelistet ist liegt schlichtweg im Verbot des Dienstes durch die chinesische Regierung. Diese versucht mit dieser Maßnahme Volksunruhen und Demonstrationen zu verhindern. Anstelle dessen gibt es in China den Twitter Klon Weibo, der unter der Kontrolle der Regierung steht und angeblich über 160 Millionen Nutzer vgl. (Hollmann, 2011) zählen soll. Sollte diese Zahl stimmen, besitzt Weibo bereits mehr User alleine in China als Twitter weltweit. Hier an genaue Zahlen zu gelangen ist jedoch durch die Zensurpolitik Chinas sehr schwer. Diese Zahlen sind daher mit Vorsicht zu behandeln. Das Indien auf Platz zwei liegt ist nicht weiter verwunderlich, wenn man sich die Bevölkerungszahl von über 1,2 Milliarden Menschen vor Augen hält. Da in Indien neben Hindi, Englisch zweite Amtssprache ist, trägt Indien sicherlich zum hohen Anteil englischer Tweets bei. Japan auf Platz drei der Platzierung deckt sich ebenfalls mit den Sprachanteilen von Twitter. Seltsam erscheint allerdings, dass es Deutsch trotz Platz vier des weltweiten Traffics nicht unter die zehn am häufigsten verwendeten Sprachen geschafft hat.

[9] Quelle: http://www.marketinggum.com/twitter-statistics-2011-updated-stats/

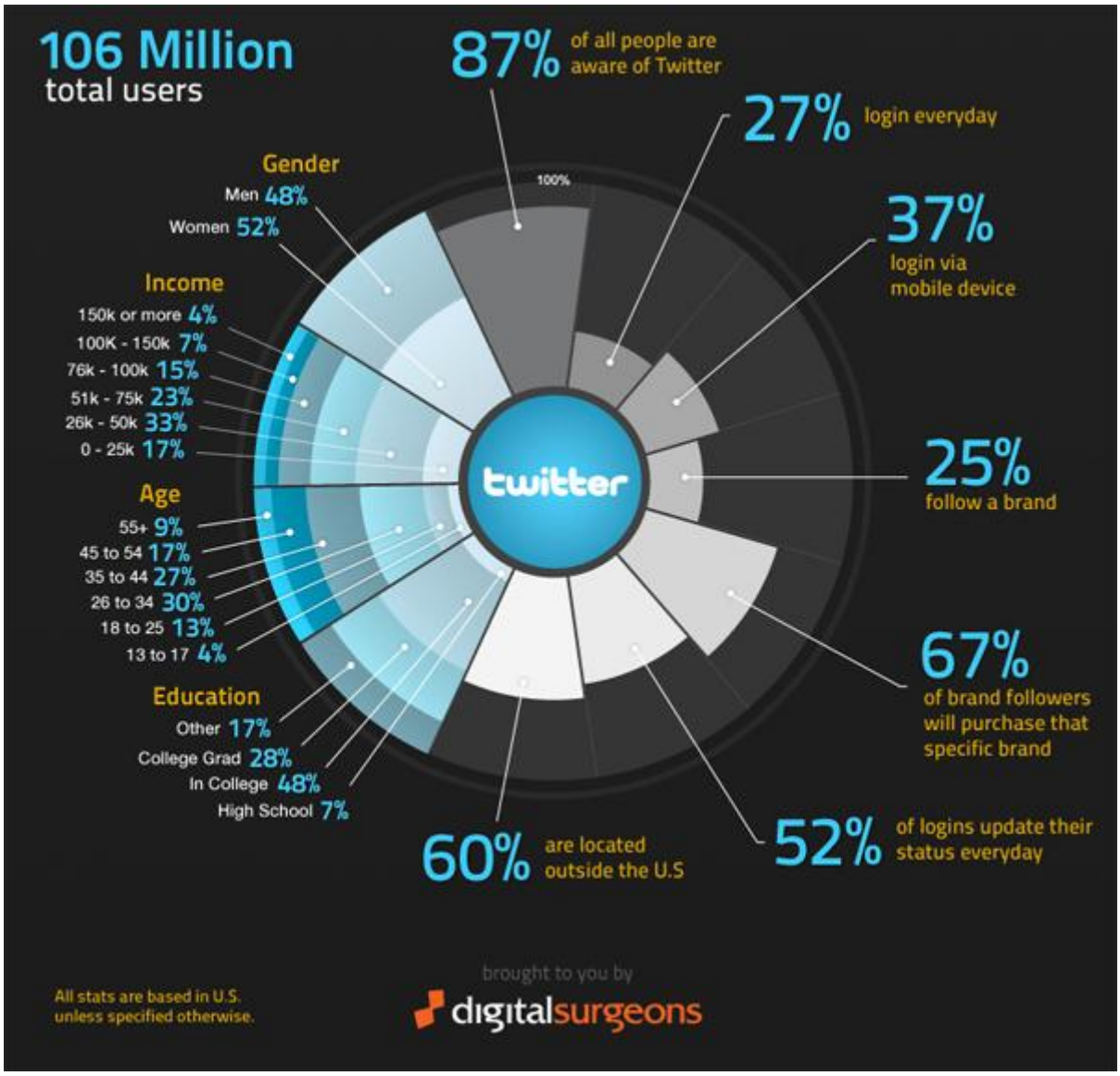

Abbildung 6: Twitter Statistik[10]

Abbildung 6 zeigt einige weitere interessante Fakten der Twitter User. Diese sollen soweit möglich mit den statistischen Daten der US Census Studien verglichen werden siehe (Census, 2009). Die Geschlechterverteilung der Twitter-User ist nahezu ausgeglichen. Der Anteil weiblicher User liegt mit 2% Vorsprung nur marginal vor dem der Männer. Die Statistik (2000) weist für die Gesamtbevölkerung eine Verteilung von Männern 49,06% und Frauen 50,94% aus. Die Einkommensverteilung zeigt eine Verbreitung von Twitter durch alle Gesellschaftsschichten. Die Census Studie von 2009 gibt hier im Vergleich für die USA folgende Einkommensverteilung der Haushalte an: 0-25k (17,8%), 25k-50k (23,9%), 50k-75k (19.3%), 75k-100k (13,5%), 100k-150k (14,9%), >150k (10,7%). Der Prozentsatz der Einkommensschicht 0-25k stimmt somit exakt mit den Daten von Twitter überein. Die Einkommensgruppe 26k-50k ist etwas überrepräsentiert, was allerdings sicherlich an der überproportionalen Beteiligung der Mittleren Altersgruppen liegen dürfte, die vornehmlich in

[10] Quelle: http://www.marketinggum.com/twitter-statistics-2011-updated-stats/

dieser Einkommensgruppe liegen. Die Einkommensgruppe 50k-75k liegt mit weniger als 4% Abweichung relativ genau an den Werten von Twitter. Die Zahlen für die Gruppe 75k-100k stimmen wieder exakt überein. Die Einkommensgruppen über 100k sind dagegen unterrepräsentiert, was wiederum an dem höheren Alter dieser Gruppe liegen könnte.

Die Altersverteilung zeigt, dass Twitter kein reines Teenager Phänomen ist sondern, dass der überwiegende Anteil der Nutzer zwischen 20 und 55 liegt. Die Census Studio der USA von 2000 ergab folgende Altersverteilung der US Bürger: 15-19 (7,18%), 20-24 (6,74%), 25-34 (14,18%), 35-44 (16,04%), 45-54(13,39%), >55(21,05%). Auffällig aber nicht ungewöhnlich ist der mehr als doppelt so hohe Anteil der Twitter User in der Gruppe 25-34 Jahre. Erstaunlich ist allerdings das die Gruppe 35-44 unterrepräsentiert ist während die nächst höhere Altersgruppe 45-54 mit einem Unterschied von nur 3% fast mit den Daten der Census Studie übereinstimmt. Die älteste Bevölkerungsgruppe ist wie zu erwarten unterrepräsentiert.

Geschäftsmodell

Nachdem Twitter sehr lange kein Geschäftsmodell vorweisen konnte und lediglich von Investoren am Leben gehalten wurde, hat das Unternehmen seit geraumer Zeit ein werbefinanziertes Geschäftsmodell eingeführt. Dies beruht auf dem Konzept so genannter „Promoted Tweets“. Hierbei werden Tweets besonders hoch eingeordnet wenn ein User nach gewissen Schlüsselwörtern sucht. Das Modell ähnelt damit Googles Adword Service. Damit nehmen Reklame-Tweets gegenüber privaten unbezahlten Konversationen eine privilegiertere Stellung ein, selbst wenn diese erheblich älter sind als andere Tweets. Neben diesen „Promoted Tweets“ gibt es „Promoted Accounts“ bei denen Firmen dafür Geld bezahlen, dass ihr Account bei Empfehlungen für neue Kanäle möglichst weit oben platziert wird. Als dritte Promotionsform hat Twitter „Promoted Trend“ eingeführt. Twitter versteht hierunter Zeit-, Kontext- und Event-sensitive Trends, die von Werbepartnern gesponsert sind. Diese bezahlten Promoted Trends erscheinen ganz oben in der Liste der Trending Topics Liste ("Trends") auf Twitter und sind als "Gesponsert" gekennzeichnet. Desweiteren lässt Twitter Werbebotschaften in den Nachrichtenstrom von Usern einfügen, auch wenn diese den Werbebetreiber nicht abonniert haben. Twitter versucht die angezeigte Werbung möglichst gut an die Vorlieben der Nutzer anzupassen. Die Werbebotschaften sollen vom Aufenthaltsortes des Users, von den Channels die er abonniert hat, sowie den Themen über die die User, denen er folgt, gerade reden abhängen.

Abbildung 7 zeigt die Steigerung der Werbeeinahmen von Twitter seit 2010 vgl. (emarketer, 2011). Man kann erkennen, dass Twitter enorme Steigerungen in seinen Einnahmen verbuchen konnte. So wurden die Einnahmen der Firma zwischen 2010 und 2011 mehr als verdreifacht. Grund für den rasant ansteigenden Umsatz dürfte die ebenso stark anwachsende User Zahl sowie die neu eingeführten Werbesysteme sein.

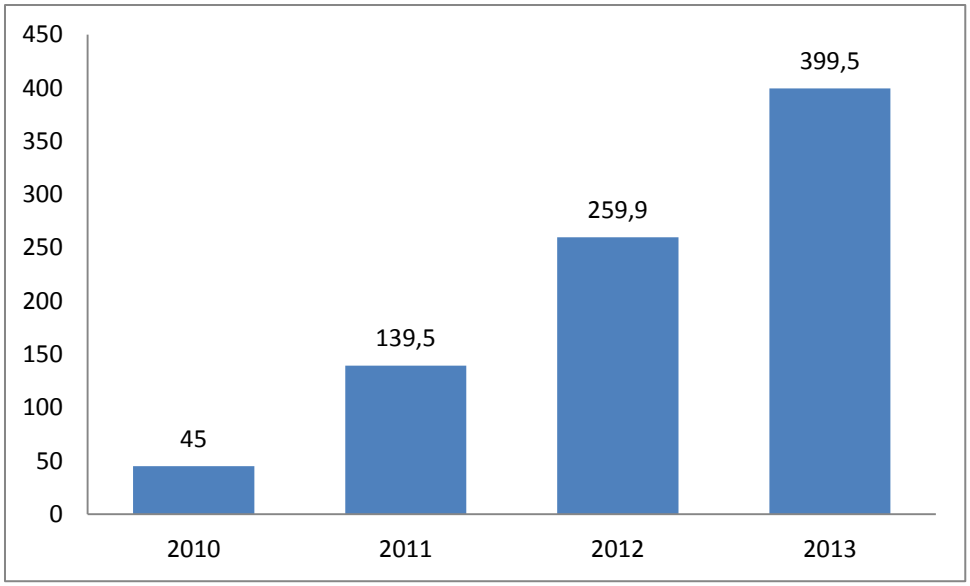

Abbildung 7: Twitter Werbeeinahmen Weltweit

3.4 Nutzercharakteristik

Verschiedene Forscher-Teams u.a. (Poblete, et al., 2011) haben untersucht, welche Unterschiede es bei der Nutzung von Twitter gibt und welche Rolle hierbei geografische Herkunft und Sprache spielen. Dabei haben sie herausgefunden, dass es Unterschiede im Maß der Aktivität, also wie viele Tweets eine Person schreibt, im Inhalt der Tweets und der Nutzung von URL´s, Re-Tweets und Hashtags gibt. Desweiteren existieren Unterschiede in der Nutzung von Twitter als Netzwerk.

Die Gründe warum Menschen Twitter nutzen sind verschieden. Um die Struktur von Twitter und damit das Nutzerverhalten besser verstehen zu können, haben einige Wissenschaftler Methoden der Netzwerkanalyse auf Twitter angewendet. Akshay Java u.a. beobachten hierbei zwei Dinge vgl. (Java, et al., 2007). Zum einen gibt es wenige Personen, die kaum Benutzern folgen während viele ihnen folgen. Zum anderen existieren Personen, die kaum Follower besitzen, während sie selbst vielen Personen folgen. Daraus leiten sie drei Kategorien von Benutzern ab: Informationsquellen, Informationssuchende und Freunde. Nach ihnen gibt es vier Intentionen zur Benutzung von Twitter: Status-Updates über die täglichen Routinen und was man gerade macht, Unterhaltungen mit der Hilfe der Antwort (Reply) Funktion. Informationsaustausch und Nachrichtenverbreitung. Viele Nutzer scheinen Twitter allerdings lediglich als Informationsquelle zu nutzen. Haewoon Kwak u.a. haben bei der Analyse der ein- und ausgehenden Verbindungen zwischen den Usern herausgefunden, dass 67,6% der Personen keine Follower besitzen, denen sie selber folgen vgl. (Kwak, et al., 2010). Es gibt also keine bidirektionale Verbindung zwischen ihnen.

Wie in jeder sozialen Struktur gibt es auch bei Twitter die Tendenz sich zu Gemeinschaften zusammen zu schließen, Informationen auszutauschen und gemeinsame Ziele zu verfolgen. De Choudhury u.a. fanden bei ihren Untersuchungen die Tendenz der User zur Homophilie also die Neigung zur Ähnlichkeit von Personen in Beziehungen vgl. (De Choudhury, et al., 2010). Nutzer mit ähnlichen Interessen tendieren also dazu sich zusammenzuschließen, beziehungsweise einander zu folgen. Der Einfluss den ein einziger User erlangen kann, hängt

nach (Meeyoung, et al., 2010) dabei nicht davon ab, wie viele Follower jemand besitzt, sondern wie viele aktive Follower er hat, die ihn zitieren und seine Nachrichten verbreiten.

Die Nutzungsweise von Twitter fällt in verschiedenen Sprachen und Kulturbereichen höchst unterschiedlich aus. Lichan Hong u.a. haben die Verwendung von URL, Hashtags, Mentions, Replies und Retweets in verschiedenen Sprachen untersucht. Folgende Tabelle zeigt deren Ergebnisse vgl. (Hong, et al., 2011).

Sprache	URL´s	Hashtags	Mentions	Replies	Retweets
Alle	21%	11%	49%	31%	13%
Englisch	25%	14%	47%	29%	13%
Japanisch	13%	5%	43%	33%	7%
Portugiesisch	13%	12%	50%	32%	12%
Indonesisch	13%	5%	72%	20%	39%
Spanisch	15%	11%	58%	39%	14%
Holländisch	17%	13%	50%	35%	11%
Koreanisch	17%	11%	73%	59%	11%
Französisch	37%	12%	48%	36%	9%
Deutsch	39%	18%	36%	25%	8%
Malaysisch	17%	5%	62%	23%	29%

Es ist sofort ersichtlich, dass die Werte zwischen den einzelnen Sprachen weit auseinandergehen. Die Deutsch und Französisch sprechenden User verwenden beispielsweise sehr viele URLs in ihren Tweets, was darauf hinweist, dass sie Twitter zu einem großen Teil zur Informationsverbreitung nutzen. Der hohe Anteil an URLs könnte auch ein Grund dafür sein, dass Deutsch unter den meist verwendeten Sprachen nicht auftaucht, da URLs möglicherweise zur Klassifikation genutzt wurden und Deutsche oft Webseiten nutzen, die ausländische Top Level Domains verwenden. Nach dem Statistik-Dienstleister Statista sind 81% aller von den Deutschen besuchten Webseiten Amerikanisch vgl. (Spiegel Online, 2012). Bei den Prozentzahlen der Mentions, also dem Erwähnen anderer User sind die Werte in allen Sprachen recht hoch. Lediglich Indonesisch und Koreanisch stechen hier hervor, da bei ihnen der Anteil von Erwähnungen über 70% liegt. Der Anteil der Replies, also dem Antworten von Usern gibt Hinweise darauf, wie sehr Twitter als Kommunikationsplattform zwischen Einzelpersonen verwendet wird. Hier liegen die Werte in den meisten Ländern auf einem ähnlichen Niveau. Nur Koreanisch hat hier mit 59% einen fast doppelt so hohen Wert als die meisten anderen Sprachen.

Lichan Hong u.a. führen weiter an, dass die am häufigsten verwendeten URLs in englischsprachigen Tweets twitpic.com, youtube.com, facebook.com, twitlonger.com und formspring.com waren vgl. (Hong, et al., 2011).

3.5 Interne Validierung

Der folgende Abschnitt wird sich mit zwei Analyseformen beschäftigen, dem automatisierten Erkennen von Sarkasmus und Glaubwürdigkeit. Dies soll sowohl die Machbarkeit dieser Analyseformen beweisen, als auch zeigen, welchen Grad an Sarkasmus und Glaubwürdigkeit Twitternachrichten besitzen.

3.5.1 Sarkasmus Erkennung

Sarkasmus ist eine der am schwierigsten zu erkennenden Eigenschaften menschlicher Sprache, nicht nur für Maschinen sondern ebenso für Menschen. Sarkasmus verdreht die Aussage eines Satzes ins Gegenteil. Für die korrekte Interpretation von Aussagen ist das Erkennen von Sarkasmus daher unabdingbar.

Eine der größten Probleme bei der Erkennung von Sarkasmus ist, dass ein Satz erst durch den Kontext in dem er geäußert wird, sarkastisch wird. Beispielsweise kann die Aussage „Das Wetter ist aber wieder schön heute." erst richtig interpretiert werden, wenn man die Wetterlage zum Zeitpunkt der Äußerung des Satzes kennt. War das Wetter zu diesem Zeitpunkt alles andere als schön, kann der Satz als sarkastisch angesehen werden. Zur richtigen Interpretation ist also ein Abgleich der Aussage mit dem Objekt beziehungsweise Sachverhalt, den er beschreibt, notwendig. Aus linguistischer und psychologischer Sicht ist Sarkasmus ein sehr gut untersuchtes Phänomen siehe ((Gibbs, 1986);(Colston & Gibbs, 2007);(Kreuz & Glucksberg, 1989);(Utsumi, 2000)). „In the context of spoken dialogues, automatic detection of sarcasm has relied primarily on speech-related cues such as laughter and prosody" (Tepperman, et al., 2006). In der Computer-Linguistik ist dies jedoch noch ein relativ junges Forschungsfeld. Die Erkennung von Sarkasmus durch einen maschinellen Algorithmus steckt daher noch in den Kinderschuhen.

Ibáñez González u.a. haben die Wirksamkeit eines maschinellen Lernansatzes zur Erkennung von Sarkasmus untersucht und wie dieser im Vergleich zu menschlicher Beurteilung abschneidet siehe (González-Ibáñez, et al., 2011). Eines der ersten Probleme auf das sie dabei stießen, war der Mangel an korrekt gelabelten Trainingsdatensätzen. Um derartige Trainingsdaten zu erlangen hat das Team Twitter-Nachrichten, in denen der Hashtag #sarcasm vorkam und damit vom Ersteller als sarkastisch gekennzeichnet waren, als Trainingsdatensätze verwendet. Dieser Vorgehensweise liegt die Idee zu Grunde, dass der Autor einer Nachricht am besten weiß, ob es sich um eine sarkastische Nachricht handelt oder nicht. Um sicher zu gehen, dass sich das Hashtag auf den ganzen Satz und nicht nur auf einen Teil davon bezieht, wurden lediglich Tweets, bei denen das Hashtag am Ende des Satzes verwendet wurde, berücksichtigt. Zusätzlich wurden als positiv und negativ gelabelte Daten mit einbezogen. Hierzu wurden wiederum Hashtags verwendet, die nahe legen, dass es sich um eine positive oder negative Aussage handelt. Beispielsweise wurden Tweets die #happy, #joy, oder #lucky enthalten als positiv angesehen und Tweets mit Inhalten wie #sadness, #angry, oder #frustrated als negativ klassifiziert.

Der gesamte Korpus dieser Analyse bestand aus jeweils 900 Tweets in den Gruppen sarkastisch, positiv und negativ. Ziel der Analyse war die Fragestellung zu beantworten, ob in Aussagen gewisse Faktoren existieren, die einen Satz als sarkastisch klassifizieren.

Bei einer manuellen Überprüfung der Trainingsdaten stellte sich heraus, dass die meisten ironischen Tweets hauptsächlich negativ gemeint jedoch als positive geschrieben waren. Der Umgekehrte Fall konnte jedoch ebenfalls beobachtet werden. Die beste Genauigkeit, die bei der Klassifikation zwischen sarkastisch, positiv und negativ erreicht werden konnte, lag bei 57%. Die höchste Genauigkeit zwischen sarkastisch und nicht sarkastisch lag bei 65%. Die Autoren weisen darauf hin, dass eine Klassifikation durch ein Lexikon oder aufgrund von Features keine brauchbaren Ergebnisse liefert. Um die Ergebnisse des Algorithmus mit der Interpretationsfähigkeit des Menschen vergleichen zu können, wurde ein Vergleichstest gestartet, bei dem drei Personen die ihnen vorliegenden Tweets in sarkastisch, positiv und negativ klassifizieren sollten. Hierzu wurden den Tweets die entsprechenden Hashtags entfernt um eine Unvoreingenommenheit zu gewährleisten. Das Ergebnis zeigte, dass alle drei Personen in lediglich 50% der Fälle mit ihrer Einschätzung übereinstimmten. Die mittlere Klassifikationsgenauigkeit lag bei 62%. Die Genauigkeit der Tweets, bei denen alle drei Personen übereinstimmten lag lediglich bei 43%. Bei einem weiteren Test, in dem die Personen lediglich zwischen Sarkastisch und nicht Sarkastisch unterscheiden sollten, wurde eine Übereinstimmung der Ergebnisse in 71% der Fälle erreicht, die Klassifikationsgenauigkeit lag nun bei 66%. Die Genauigkeit bei den Tweets, in denen alle übereinstimmten, lag in diesem Fall mit 59% etwas höher. Wenn man die Ergebnisse der Untersuchung zwischen Mensch und Maschine miteinander vergleicht, stellt man fest, dass die Klassifikationsgenauigkeit nahezu identisch ist. Im Test zwischen sarkastisch, positiv und negativ liegt der Mensch mit 62% nur 5% über der Maschine. Im Test zwischen sarkastisch und nicht sarkastisch trennen beide nur 1%.

In einem weiteren Test sollte überprüft werden, welche Rolle sprachunabhängige Klassifikationsmerkmale spielen. Hierzu wurden lediglich Tweets berücksichtigt die Emoticons[11] enthalten. Zwei Personen erreichten bei der Klassifikation zwischen sarkastisch und nicht sarkastisch eine Übereinstimmung von 89% und eine Klassifikationsgenauigkeit von je 73%. In den 89%, in denen beide übereinstimmten, lag die Genauigkeit jetzt bei 70%. Der Algorithmus erreicht mit 71% eine wiederum nahezu identische Genauigkeit.

Es lässt sich also erkennen, dass sprachunabhängige Merkmale wie Emoticons die Einschätzbarkeit von Sarkasmus erheblich erhöhen. Die Entscheidungsträger gaben an, die Aufgabe als schwierig empfunden zu haben. Hauptgründe waren hierfür die Kürze der Nachricht und der Mangel an Kontext-Informationen. Eine der Personen gab an, dass es gelegentlich notwendig war Allgemeinwissen über aktuelle Themen mit einfließen zu lassen. Für eine automatisierte Erkennung müsste ein Algorithmus derartige Informationen mit einbinden. Desweiteren ist es notwendig bei Tweets, die Antworten auf andere Tweets darstellen, die Vorgeschichte der Unterhaltung zu kennen.

Die Ergebnisse dieser Studie zeigen, dass Menschen bei der Erkennung von Sarkasmus nicht besser abschneiden als der hier vorgestellte Algorithmus.

[11] Emoticon ist eine Wortkreuzung aus Emotion und Icon. Als Emoticons wird eine Zeichenkette bezeichnet, die einen gefühlsbezogen Gesichtsausdruck nachempfunden ist (Bsp. „:-)“, „;-]“). Hierdurch sollen gewisse Gefühle bzw. Emotionszustände ausgedrückt werden.

3.5.2 Glaubwürdigkeitserkennung

Bevor wir mit der Analyse von Twitter-Nachrichten beginnen, sollten wir überprüfen, als wie glaubwürdig die User Nachrichten von Twitter eigentlich ansehen. Falls der Großteil der User die ihnen zukommenden Nachrichten als vollkommen unglaubwürdig betrachtet und ihnen daher keine Beachtung schenkte, wäre der Versuch eine Vorhersage aufgrund dieser Daten zu tätigen sinnlos.

Am Anfang der Verbreitung von Twitter wurde die Informationsgüte beziehungsweise der Informationsinhalt der Tweets von vielen als extrem niedrig eingeschätzt. Durch die zunehmende Anzahl an Personen aus Wissenschaft und Politik bei Twitter hat sich dieser Umstand sicherlich gebessert, jedoch bis zu welchem Grad? Carlos Castillio u.a. haben die Glaubwürdigkeit von Twitter Nachrichten untersucht um herauszufinden, ob es Indikatoren in den Nachrichten gibt, mit denen sich die Glaubwürdigkeit messen lässt vgl. (Castillo, et al., 2011). Ihre Hypothese war, dass es in der sozialen Netzwerkstruktur von Twitter Signale geben muss, mit denen sich auf die Glaubwürdigkeit schließen lässt. Ihrer Aussage zu folge, entspricht der meiste Teil der Nachrichten zwar der Wahrheit, allerdings würde Twitter auch dafür genutzt um Falschinformationen und Gerüchte zu verbreiten. Dies sogar sehr oft, ohne dass die User dies beabsichtigen. Hierbei definieren Castillo u.a. Glaubwürdigkeit als die subjektive Wahrnehmung der User auf Richtigkeit. Ob die Aussage tatsächlich korrekt ist, wurde nicht berücksichtigt. Es soll also untersucht werden, durch welche Indikatoren ein User eine Nachricht als glaubwürdig betrachtet oder nicht. Der Focus wurde hier auf zeitkritische Informationen über aktuelle Ereignisse gelegt. Um eine Menge an derartigen Themen zu bekommen, deren Glaubwürdigkeit getestet werden kann, wurde eine Testgruppe mit der Aufgabe betraut, Twitter Nachrichten nach nennenswerten Nachrichten zu durchsuchen und zu klassifizieren. Für die folgenden Tests wurden dann diejenigen Themen aus den als Nachrichten klassifizierten Tweets herausgesucht, die als von allgemeinem Interesse angesehen werden können und nicht nur für eine kleine Nutzergruppe gedacht sind. Die hierdurch gewonnen Nachrichten-Tweets wurden einer zweiten Gruppe vorgelegt, die diese als wahr oder falsch klassifizieren sollte. Um die nötige Anzahl an Nutzer zu gewinnen, haben die Forscher Amazons Mechanical-Turk[12] Service genutzt.

Der Anteil an als relevant angesehenen Tweets lag bei 28,5%. Aus dieser Menge wurden zufällig 383 Themen ausgewählt, deren Glaubwürdigkeit überprüft werden sollte. In der folgenden Analyse wurden vier Grade von Glaubwürdigkeit unterschieden: „almost certainly true“, „likely to be false“, „almost certainly false“, und “I can’t decide”. Im Usertest mit dieser Gliederung bewerteten fast alle Personen die Nachrichten als „likely to be true”, was für eine eindeutige Klassifikation nicht ausreicht. In einem zweiten Test wurde daher diese Option aus den Auswahlkriterien herausgenommen. Dieser Test kam zu folgendem Ergebnis: almost certainly true (41%), likely to be false (31.8%), almost certainly false (8.6%) und ambiguous (18.6%). Ein Algorithmus, der die Usereinteilungen für Nachrichten und als Trainingsdaten verwendete, kam auf eine Klassifikationsgenauigkeit von 89%. Er kann damit als relativ zuverlässig angesehen werden. Im Folgenden wurden die Trainingsdaten der

[12] Amazons Mechanical Turk ist eine Internet Crowdsourcing Plattform bei der Programmierer Aufgaben, die maschinell nur schwer oder unzureichend genau zu erledigen sind, von menschlichen Entscheidungsträgern bearbeiten lassen können.

Usereinschätzung für den Glaubwürdigkeitsgrad als Trainingsdatensatz verwendet. Der Algorithmus sollte hier zwischen „almost certainly true“ gegenüber allen anderen Einstufungen mit Ausnahme der „ambiguous“ Einstufung unterscheiden. Diese wurde hierbei weggelassen. Bei diesem Test konnte eine Klassifikationsgenauigkeit von 86% Prozent erreicht werden.

Um festzustellen welche Eigenschaften einen Tweet besonders glaubwürdig erscheinen lassen, wurde mit den Klassifikationsdaten ein Entscheidungsbaum erstellt. In diesem konnte beobachtet werden, dass Tweets die keine URL enthalten tendenziell unglaubwürdiger sind. Negative Sentimente lassen einen Tweet dagegen glaubwürdiger, positive Sentimente unglaubwürdiger erscheinen. Desweiteren werden User, die in der Vergangenheit bereits sehr viele Nachrichten geschrieben haben oder sehr viele Freunde besitzen als glaubwürdiger angesehen. Ein weiteres positives Indiz auf Glaubwürdigkeit war die Anzahl an Retweets, die eine Nachricht bekommen hat. Diese Ergebnisse könnte man dahingehend interpretieren, dass Tweets, die Links enthalten glaubwürdiger sind, weil ihre Urheber damit Quellen für ihre Aussagen angeben und ihre Aussage somit überprüfbar machen. Dass die Anzahl der bereits geposteten Tweets sich positiv auf die Glaubwürdigkeit auswirkt, könnte daran liegen, dass Personen generell anderen Personen mehr trauen, wenn sie diese schon eine Zeit lang kennen und mit deren Nachrichten bereits gute Erfahrungen gesammelt haben. Jemand der viele Freunde hat, macht naturgemäß einen glaubwürdigeren Eindruck, da er, um sich die Freundschaft anderer nicht zu verspielen, in der Regel keine Fehlinformationen verbreitet. Die Anzahl an Retweets weist desweiteren darauf hin, dass andere User die Äußerungen einer Person als so bedeutsam ansehen, dass sie diese an ihren Followerkreis weiterleiten. Dass diese Nachrichten als besonders glaubwürdig gelten, verwundert daher nicht.

Das generelle Vertrauen der User gegenüber Online-Nachrichten scheint ebenfalls recht gut zu sein. Nach (Flanagin & Metzger, 2000) vertrauen die User Online-Nachrichten ebenso wie den meisten traditionellen Medien. Einzige Ausnahme bilden hier Zeitungen, die als generell glaubwürdiger angesehen werden. Nach einer US-Studie (Pew Research Center, 2008) ist das Internet desweiteren die zweitwichtigste Quelle für Nachrichten nach Fernsehen bei Personen unter 30 Jahren. Internetnutzer machen aber durchaus einen Unterschied zwischen den Nachrichtenquellen im Netz. Nachrichtenseiten werden beispielsweise als weitaus vertrauenswürdiger angesehen als Blogs vgl. (Princeton Survey Research, 2005). Hauptgrund hierfür dürfte die Tatsache sein, dass Blogs größtenteils Inhalt von Privatpersonen enthalten und nicht von professionellen Journalisten erstellt wurden. Twitter wird jedoch durchaus von Personen aus der Nachrichtenbranche verwendet um Neuigkeiten möglichst schnell publizieren zu können. Andrew Flanagin und Miriam Metzger haben herausgefunden, dass Personen im Online Umfeld sich sehr leicht von der äußeren Aufmachung einer Nachricht beeinflussen lassen vgl. (Flanagin & Metzger, 2000). Dabei würden sie sich stark am Design orientieren, auch wenn dieses nichts mit dem Inhalt einer Nachricht zu tun hat. Sie haben in einem Experiment dieselbe Nachricht in unterschiedlichen Formaten einem Kreis von Testpersonen vorgelegt und nach der Glaubwürdigkeit der Nachricht gefragt. Das Ergebnis war, dass die Personen der Nachricht mehr Glauben schenkten, wenn sie als traditionelle Medienseite präsentiert wurde. Wenn dieselbe Nachricht den Personen als Blogeintrag oder Twitterpost vorgelegt wurde, haben sie diese als weit weniger verlässlich eingestuft. Trotz

dieser skeptischen Haltungen scheinen die User von Twitter den Dienst insgesamt als durchaus vertrauenswürdig einzuschätzen. Vor allem die hohen Glaubwürdigkeitseinschätzungen in der Untersuchung von (Castillo, et al., 2011) weisen hierauf hin. Twitter-Nachrichten können somit durchaus andere User überzeugen und hierdurch auch beeinflussen. Die Analyse dieser Daten macht daher durchaus Sinn wie einige folgende Beispiele beweisen werden.

3.6 Externe Validierung

Kommen wir nun zu einigen Anwendungsbeispielen für die Analyse von Twitterdaten. Diese sollen zeigen, ob ein Zusammenhang zwischen den Daten in Twitter und Daten aus anderen statistischen und wissenschaftlichen Quellen besteht. Die Ergebnisse werden zeigen, ob es Sinn macht, ein gesellschaftliches Phänomen mit Hilfe von Twitter beschreiben zu wollen.

3.6.1 Realweltsensor

Da Twitter-User in ihren Nachrichten schreiben was sie gerade tun, fühlen, hören und sehen, stellt sich die Frage, ob Twitter als Sensor für reale Phänomene dienen kann. Tetsuro Takahashi u.a. haben versucht diese Frage zu beantworten. Hierzu haben sie die Ausbreitung von Heuschnupfen mit der Verbreitung von Pollen in Japan verglichen siehe (Takahashi, et al., 2011). Ziel war es eine Heuschnupfenkarte zu erstellen, ähnlich eines Wetterreports. Twitter wurde dazu genutzt, um die Ausbreitung der Krankheit zu messen. Die Daten zur Ausbreitung der Pollen wurden durch das Pollen Überwachungssystem des japanischen Umweltministeriums erstellt. Jede Präfektur in Japan hat durchschnittlich 3,1 dieser Überwachungsstationen. Eine Herausforderung stellte dabei die Geolokalisierung der Twitter Nachrichten dar. Die Twitter API ist zwar in der Lage Längen und Breitengrade zu speichern, allerdings nutzen dieses Feature nur sehr wenige User. In den Datensätzen dieser Analyse beinhalteten lediglich 0,6% aller Nachrichten entsprechende Geodaten. Aus diesem Grund wurde versucht, aus den Standortangaben des Userprofils die Position des Users herzuleiten. Diese gibt jedoch nur den angegebenen Heimatort des Users und nicht seine aktuelle Position an. Um die Position des Heimatortes zu bestimmen, wurden die Daten mit einem Lexikon aller Orte und Präfekturen in Japan verglichen. Nach der Geolokalisierung wurde versucht, die Ausbreitung des Heuschnupfens durch die Twitter Nachrichten „ich bekomme gerade Heuschnupfen“ oder „ich bekomme keinen Heuschnupfen“ zu klassifizieren. Die Genauigkeit des Klassifikationsalgorithmus lag bei 77,27%. Das Resultat dieser Analyse war, dass Twitter- Nachrichten durchaus als Sensor für Reale Phänomene dienen können. Dabei wurde die Übereinstimmung zwischen Twitter und Messstationen umso höher, je mehr Tweets in die Analyse mit eingebunden wurden.

3.6.2 Vorhersagbarkeit von Wahlen

Eine der ersten Versuche das Verhalten von Personen mit der Hilfe einer Twitteranalyse vorherzusagen, war der Versuch Twitter als Umfragebasis für politische Wahlen zu nutzen. Daher gibt es bereits verschiedene Untersuchungen auf diesem Gebiet in unterschiedlichen Ländern. Deren Ergebnisse was die Verlässlichkeit derartiger Untersuchungen anbelangt, gehen dabei jedoch sehr auseinander. Demgemäß soll hier eine differenzierte Erläuterung deren Ergebnisse und eine Kritik an ihrem Vorgehen erfolgen. Der Grund für das große Interesse an Twitter im Kontext politischer Wahlen liegt in der Tatsache, dass Twitter einen Echtzeitstream einer großen Menschenmenge liefern kann. Umfragen zu Politikern und

Parteien dauern dagegen immer eine gewisse Zeit, bis sie Ergebnisse liefern können. Da sich die öffentliche Meinung zu einem Politiker oder einer Partei jedoch sehr schnell ändern kann, könnte ein Echtzeitstream wie Twitter hier zu zeitlich besser zurechenbaren Ergebnissen führen.

Als wichtiger Startpunkt sozialer Medien im politischen Kontext kann wohl der US-Wahlkampf 2008/2009 angesehen werden, in dem Barack Obama als erster Präsidentschaftskandidat der USA massiv soziale Netzwerke wie Facebook und Twitter als Kampagnenplattform genutzt hat. Dieses Ereignis hat den Beweis erbracht, welchen Beitrag soziale Netzwerke in einem Wahlkampf leisten können.

Inwiefern werden diese Plattformen jedoch für einen politischen Diskurs und somit für einen Meinungsaustausch genutzt? Andranik Tumasjan u.a. haben versucht genau diese Frage zu beantworten um damit festzustellen, ob Twitter ein verlässliches Bild der öffentlich politischen Meinung sein kann vgl. (Tumasjan, et al., 2010). Wie genau ist jedoch dieses Bild, und kann damit das Ergebnis einer Wahl vorhergesagt werden? Hierzu haben die Wissenschaftler 104.003 Tweets in den Wochen vor der Bundestagswahl am 27 September 2009 gesammelt. Hierbei standen die Forscher vor dem Problem, dass Twitter-Nachrichten durch ihre Länge viel weniger Informationen enthalten, als Artikel in Blogs oder Zeitungen. Nach einer Marketing Consultingfirma enthalten desweiteren 40% des Twitter Traffics nur sinnloses Gerede vgl. (PearAnalytics, 2009). Dazu kommt, dass sehr viele Nachrichten URLs enthalten und somit ein Teil der Information nicht in der Nachricht selbst, sondern auf der verlinkten Webseite liegt. Zu guter Letzt ist die Nachrichtenverteilung über die Nutzer ungleichmäßig verteilt. Eine kleine Menge an Usern ist somit für eine große Menge an Nachrichten verantwortlich. Trotz dieser Probleme können nach ihrer Meinung Twitter-Nachrichten durchaus sinnvolle Aussagen über Parteien und Politiker enthalten. „these messages illustrate that tweets can contain a lot of relevant information. So despite their brevity substantive issues can be expressed in 140 characters or less" (Tumasjan, et al., 2010). Für die Auswertung der Analyse maßen die Forscher die prozentuale Verteilung der Tweets zu den einzelnen Parteien und berechneten die Nähe der einzelnen Parteien durch das gemeinsame Auftreten in einer einzelnen Nachricht.

Die Ergebnisse der Verteilung sind in folgender Tabelle enthalten. Man kann erkennen, dass die Verteilung der Tweets relativ exakt mit der späteren Stimmenverteilung übereinstimmt.

Partei	Anteil am Traffic	Wahlergebnis	Vorhersagefehler
CDU	30,10%	29%	1%
CSU	5,60%	6,90%	1,30%
SPD	26,60%	24,50%	2,20%
FDP	17,30%	15,50%	1,70%
LINKE	12,40%	12,70%	0,30%
Grüne	8%	11,40%	3,30%
		MAE:	1,65%

Damit kann Twitter eine fast gleichwertige Vorhersage über den Ausgang einer Wahl treffen wie die meisten etablierten Meinungsforschungsinstitute wie folgende Tabelle zeigt.

Quelle:	MAE
Twitter	1,65%
Forsa	0,84%
Allensbach	0,80%
Emnid	1,04%
Forschungsgruppe Wahlen	1,48%
GMS	1,40%
Infratest/dimap	1,28%

Die Ergebnisse der Mengenverteilung decken sich mit den Ergebnissen einiger weiterer Forscher. Dabei soll die Anzahl an Erwähnungen eines Politikers in der Presse das Wahlergebnis besser vorhersagen können als Meinungsstudien. „In sum, the joint mentions of political parties accurately reflect the political ties between the parties“ (Tumasjan, et al., 2010).

Während diese Untersuchung zu vielversprechenden Ergebnissen gelangte, äußerten andere wissenschaftliche Quellen Kritik an der Machbarkeit einer solchen Analyse. Panagiotis Metaxas u.a. überprüften die Ergebnisse mit dem Versuch die Wahl eines US Senators vorherzusagen vgl. (Metaxas, et al., 2011). Ihre Ergebnisse sind in folgender Tabelle zu sehen. Hierzu verwendeten sie die prozentualen Anteile der Kandidaten, an der Gesamt-Twitter-Nachrichtenmenge.

	Coakley		Brown	
	Tweets	%	Tweets	%
Pre - Wahl	52116	53,86%	44654	46,14%
Wahltag	21076	49,94%	21123	50,06%
Post - Wahl	14381	29,74%	33979	70,26%
Total	87573	46,75%	99756	53,25%

Während des Wahltages stimmen die Ergebnisse relativ gut überein (Brown gewann mit 51.9% der Stimmen). Vor und nach dem Wahltag gehen sie allerdings erheblich auseinander. Hiermit wäre das Vorhersageergebnis fehlerhaft gewesen. Nach den Tweet-Anteilen vor dem Wahltag wäre hier Coakley und nicht Brown als Sieger vorhergesagt worden. Die Twitter Volumen Methode hatte vor der Wahl dabei eine Fehlerrate von 5,76%. Dieselbe Wahl wurde von den Forschern danach als Sentiment-Analyse durchgeführt. Es wurde also nicht nur die Menge, sondern auch die Meinung einer Person zu einem Politiker mit einbezogen. Diese Methode sagte Brown korrekt als Sieger hervor. Vor der Wahl hatte die Sentiment Methode eine Fehlerrate von gerade einmal 1,1%. Die Autoren stehen dieser Analyseform jedoch sehr kritisch gegenüber.

In einem weiteren Test analysierten sie die Korrektheit der Einstufung des Algorithmus in die Ausprägungen positiv, negativ und neutral. Der Algorithmus arbeitete hierbei mit einem

Wörterbuch um die Tweets zuzuordnen. Um die Korrektheit zu überprüfen wurde eine Menge von Tweets manuell klassifiziert und diese Einstufung mit der des Computers verglichen. Die Menge an korrekt eingestuften Tweets lag gerade einmal bei 36,85% und somit nur knapp über einer zufälligen Verteilung von 33% in den drei Ausprägungen. Der Algorithmus ist somit ganz klar nicht in der Lage die Intention einer Aussage herauszufinden, beziehungsweise die politische Präferenz einer Person zu bestimmen. Nach ihrer Aussage sind die derzeitigen Analysemethoden damit nicht besser als eine zufällige Schätzung. „we find that electoral predictions using the published research methods on Twitter data are not better than chance" (Metaxas, et al., 2011). Beide Analysemethoden konnten den Gewinner der Wahl in lediglich der Hälfte der Fälle bestimmen. Die Forscher kritisieren weiterhin, dass die Altersstruktur der Twitter User nicht mit dem der Wähler übereinstimmen würde. Somit kämen die Aussagen zu den Politikern in Twitter von anderen Personen als die Stimmen bei der Wahl.

Wie wir bei der Altersstruktur im vorherigen Kapitel gesehen haben, sind vor allem ältere Menschen in Twitter unterrepräsentiert. Die korrekte Auswahl der Testpersonen ist jedoch eine der wichtigsten Faktoren bei der Meinungsanalyse von Wahlen. „the most important aspect of correct prediction is the selection of a representative and unbiased sample of the population" (Metaxas, et al., 2011). Weiterhin wurde kritisiert, dass viele Forscher soziale Medien als Blackbox verwenden, die ihnen eine Antwort liefert, aber keine Begründung wie es zu dieser Antwort kam. Die Autoren geben an, dass der Algorithmus die Unterschiede zwischen Sozialen-Medien-Daten und Daten von natürlichen Phänomenen mit einbeziehen muss. Desweiteren muss klar sein, warum ein Algorithmus funktioniert.

Die Untersuchung im Umfeld politischer Wahlen hat gezeigt, dass die Menge an Twitter-Nachrichten durchaus mit der Verteilung menschlicher Entscheidung übereinstimmen kann. Es stellt sich somit die Frage, ob diese ebenfalls auf die Kaufentscheidung am Aktienmarkt zutrifft. Ob eine Sentimentanalyse dabei zu besseren Ergebnissen führt als die Messung der Nachrichtenmenge bleibt jedoch weiterhin offen, da in der Untersuchung von Metaxas kein funktionierender Sentimenterkennungsalgorithmus verwendet wurde.

3.6.3 Vorhersage von Börsenkursen

In diesem Abschnitt soll die derzeitige Forschungslage bei der Vorhersage von Börsenkursen mit Hilfe von Twitter Nachrichten erläutert werden. Da diese Analyseform auch den Inhalt der späteren Untersuchungen bildet, soll hierauf besonders detailliert eingegangen werden.

Devendra Taytal und Satya Komaragiri haben untersucht, ob ein Zusammenhang zwischen der Meinung von Twitter- und Blog-Nutzern und dem Aktienkurs von Unternehmen besteht vgl. (Taytal & Komaragiri, 2009). Dieser Untersuchung lag die Idee zu Grunde, dass die Wahrnehmung der Öffentlichkeit zu einer Firma oder einem Produkt zu einem großen Teil dazu beiträgt, wie sich deren Güter oder Dienstleistungen verkaufen. Dies legt nahe, dass die öffentliche Meinung, die bei der Analyse von Blogs- und Twitter-Nachrichten gesammelt werden kann, einen Ausblick auf die zukünftigen Verkaufszahlen eines Unternehmens geben könnte. Für diese Untersuchung sammelten die Forscher über einen Zeitraum von einem Monat mit der Hilfe der Twitter API alle Tweets sowie mit Hilfe eines Web Crawlers eine Fülle von Blogs und den Aktienkurs der Firmen Google und Microsoft. Ziel war es den

Aktienkurs des nächsten Tages vorhersagen zu können. Zur Vorhersage des Aktienkurses des nächsten Tages wurde der Aktienwert des Vortages als Basiswert genommen und eine Durchschnittsmeinungsausprägung der Nachrichten des aktuellen Tages als Delta auf diesen Vergangenheitswert addiert. Hierbei wurde eine Meinungsskalierung von -5 für besonders schlecht bis +5 für besonders gut verwendet. Die exakte Additionsmenge auf den Aktienwert wurde durch eine einen Monat dauernde Trainingsphase ermittelt. Hierzu wurde der durch die Analyse vorhergesagte Wert immer wieder mit dem realen Aktienkurs verglichen und die Gewichtung des Meinungsdelta entsprechend angepasst. Je länger dieser Vorgang wiederholt wurde, desto genauer fiel die Vorhersage aus. Die nötigen Daten hierzu wurden einen Monat lang in einer Datenbank gesammelt. Für die Trainingsphase wurde somit versucht die bereits bekannten Aktienkurse vorherzusagen. Die Testphase wurde über den Zeitraum von 2 Monaten durchgeführt. Der hier entwickelte Algorithmus konnte im Falle der Twitter Nachrichten eine Vorhersagegenauigkeit von 97,2% (Google) und 91,1% (Microsoft) erreichen. Die Ergebnisse der Blog-Analyse lagen bei 91,1% (Goolge) und 76,6% (Microsoft). Im Vergleich schnitt die Microblogging Analyse somit wesentlich besser ab als die von regulären Blogs. Die Forscher machen hierfür verschiedene Gründe verantwortlich. Nach ihnen liegt der Hauptgrund in der besseren Klassifizierbarkeit von Twitter Nachrichten. Die reduzierte Länge derartiger Nachrichten vereinfacht die Zuordbarkeit durch einen Algorithmus. Die meist aus einem Satz bestehenden Tweets, werden als wesentlich repräsentativer für die Intention einer Person angesehen, als ein ganzer Text. Als weiteren Grund sehen die Autoren die soziale Netzwerkstruktur von Twitter an. Durch Retweets können sich Nachrichten exponentiell ausbreiten und die Möglichkeit auf Nachrichten zu antworten bildet die Basis für Konversationen. Devendra Taytal und Satya Komaragiri geben eine positive Prognose für die weitere Nutzung dieser Analysemethode vgl. (Taytal & Komaragiri, 2009). Sie gehen davon aus, dass durch die steigende Qualität der Text Mining und Sentiment-Analysemethoden die Ergebnisse von Vorhersagen auf diesem Gebiet immer besser werden.

Die psychologische Forschung sagt uns, dass Emotionen zusammen mit Informationen einen großen Einfluss auf menschliche Entscheidungen haben vgl.(Damasio, 1994), (Kahneman & Tversky, 1979). Auch wirtschaftliche Entscheidungen werden von Gefühlszuständen stark beeinflusst. „social mood affects the decisions of consumers, investors, and corporate managers alike“. „…stock market itself is a measure of social mood“ (Nofsinger, 2005). Johan Bollen u.a. haben untersucht, ob dies auch auf die Gesellschaft als Ganzes zutrifft vgl. (Bollen, et al., 2011). Anders ausgedrückt, gibt es emotionale Zustände die das kollektive Handeln beeinflussen. Hierzu verglichen die Forscher die Werte des Dow Jones Industrial Average (DJIA) Index mit Nachrichten aus Twitter. Es wurden dabei zwei Analyse-Werkzeuge genutzt, Opinion- Finder und Googles Profile of Mood States (GPOMS). Opinion Finder wurde dazu genutzt, um positive und negative Aussagen zu identifizieren. GPOMS versucht dagegen aus den Nachrichten eine von sechs Gefühlszuständen zu messen. Die Autoren geben an, dass der Aktienkurs am meisten von Nachrichten beeinflusst wird und weniger von aktuellen oder vergangenen Preisen. Ziel der Untersuchung war es daher herauszufinden, ob die öffentliche Stimmung einen Aktienkurs genauso beeinflusst wie aktuelle Nachrichten. Dabei sollte nicht ein kausaler Zusammenhang bewiesen werden, sondern ob eine Zeitlinie eine Vorhersagekraft auf eine andere hat. Der Inhalt eines Tweets ist

zwar nicht besonders aussagekräftig. Die Analyse von Millionen von Tweets kann jedoch eine recht gute Aussage über die gesellschaftliche Stimmung liefern. Bei einem Vergleich der Kurven von Opinion Finder und GPOMS gab es Ähnlichkeiten zwischen dem Fröhlichkeitsindex des GPOMS Systems und der Kurve des Opinion Finders. Als Gegentest wurde ein Vergleich der Stimmungslage mit verschiedenen kulturellen Ereignissen gestartet, mit dem Ergebnis, dass die GPOMS Werte sich mit einigen überlagerten. Gewisse Gefühlsausprägungen korrelieren dabei mehr, manche weniger mit Ereignissen. Das Ergebnis dieser Untersuchung war, dass die Opinion-Finder- Analyse keinerlei Vorhersagekraft auf den DJIA hatte. Die Calm Dimension des GPOMS konnte dagegen eine gewisse Korrelation zeigen. Alle anderen Dimensionen konnten dagegen keine besseren Ergebnisse liefern, als die des Opinion-Finders. Desweiteren konnte die Dimension „Happy“ für sich alleine zwar keine Vorhersagekraft beweisen, die Autoren weisen jedoch darauf hin, dass diese Dimension die Vorhersagekraft verbessern könnte, wenn diese mit Calm kombiniert wird. Der Calm-Indikator konnte dabei Kurven vorhersagen, die erst drei bis vier Tage später im DJIA erschienen. Als finales Ergebnis dieser Analyse gaben die Autoren an, dass die Standardmodelle der Aktienkursvorhersage durch eine Einbeziehung von emotionalen Messungen signifikant verbessert werden können.

Nach der Efficient Market Hypothesis (EMH) sind Finanz Märkte dabei informell effizient. Dies bedeutet, dass der Marktpreis alle bekannten Informationen wiederspiegelt. Der Marktpreis richtet sich also nach der Menge öffentlich verfügbarer Informationen und passt sich spontan an neue Informationen an. Dies bedeutet, dass Investoren keinen Profit erwirtschaften können, wenn sie die für ihre Handelsstrategien öffentlich verfügbaren Informationen nutzen vgl. (Fama (1970), Fama (1991)). Neuere Studien belegen jedoch, dass der Markt nicht immer der aktuellen Nachrichtenlage folgt vgl. (Malkiel (2003)). Nachrichten werden demnach vom Marktpreis nicht sofort und vollständig aufgenommen. Nach Bagnoli, Beneish, and Watts (1999), sollen beispielsweise Gerüchte, die unter Tradern kursieren, die Erwartungshaltung des Marktes besser beschreiben. Es existieren eine Vielzahl von Foren und Blogs in denen Trader derartige Gerüchte untereinander austauschen. Vor allem Daytrader nutzen solche Plattformen vgl. (Koski, et al., 2004). Eine dieser Plattformen ist Twitter. Timm Sprenger und Isabell Welpe haben nun untersucht, ob sich die Ansichten von Tradern in Twitter auf Aktienkurse auswirken vgl. (Sprenger & Welpe, 2010). Dabei haben sie versucht, aus den Tweets gewisse Kauf-, Verkaufs- und Haltesignale zu erkennen, folglich ob jemand empfiehlt eine Aktie zu kaufen, zu verkaufen oder sie zu halten. Hierzu wurde ein Klassifikationsalgorithmus verwendet, der diese Signale im Tweettext erkennen sollte. Dieser Algorithmus konnte in der Trainingsphase eine Genauigkeit von 81,2% erreichen. Die Motivation hinter ihrer Analyse war die Ansicht der Autoren, dass die bisherigen Twitter-Analysen die Finanzmarkt-Indikatoren eines Markt- Modells nur unzureichend implementiert haben. Desweiteren hatten ihrer Meinung nach bisherige Studien der Finanz Community das Problem, eine zu kleine Teilmenge aller User- Meinungen mit einzubinden. Eines der Erkenntnisse dieser Untersuchung war, dass die Anzahl an aktienrelevanten Tweets mit der Anzahl an gehandelten Aktien anstieg, was auf eine Korrelation zwischen den beiden Werten hinweist. Dabei konnte die Nachrichtenmenge des letzten und vorletzten Tages das Handelsvolumen des aktuellen Tages vorhersagen. Schwankende Kurse haben ebenfalls zu einer größeren Nachrichtenmenge geführt. Dies weist darauf hin, dass bei einer hohen

Unsicherheit die Trader mehr Informationen austauschen. Es konnte jedoch kein Zusammenhang zwischen der Nachrichtenmenge und den Handelsgewinnen festgestellt werden. Ein weiteres Ergebnis war, dass trotz der hohen Klassifikationsgüte des Algorithmus, die aggregierte Nachrichtenanalyse keine klaren Kauf- und Verkaufssignale erkennen konnte. Kaufsignale bei einer Aktie resultierten allerdings in überdurchschnittlichen Gewinnen. Bei Verkaufssignalen konnte dieser Zusammenhang jedoch nicht hergestellt werden. Ihre Untersuchungen zeigten ebenfalls, dass neue Informationen in den Tweets sehr schnell vom Marktpreis aufgenommen werden, wodurch es schwierig wird diese Informationen auszunutzen. „However, our results indicate that new information, reflected in the tweets, is incorporated in market prices quickly, and reasonable transaction costs make it difficult to exploit market inefficiencies" (Sprenger & Welpe, 2010). Eine Untersuchung des Userverhaltens ergab, dass Usern, die überdurchschnittliche Investmentratschläge gaben, eine höhere Glaubwürdigkeit gegeben wurde. Diese äußerte sich in Twitter durch eine höhere Anzahl an Re-Tweets und durch eine größere Anzahl an Followern. Die Analyse einzelner Tweets ergab jedoch, dass höherwertige Informationen nicht zwangsläufig in einer größeren Anzahl an Re-Tweets und Followern münden. Die Gewinnspanne die Sprenger und Welpe durch diese Analyseform erreichen konnten, schwankte je nach Firma sehr stark. Von -56,3% bei Monsanto bis hin zu +22,1% für Regions Financial Corp. Gewisse Handelsstrategien haben sich dabei als lukrativer als andere erwiesen. Vor allem die Nutzung von Twitter-Nachrichten des vergangenen Tages und das Halten von Aktien für acht Tage konnten gute Ergebnisse liefern.

4 Sentimentanalyse

Das folgende Kapitel wird die Kernanalyseform der späteren Untersuchung beschreiben, die Sentimentanalyse. Diese ist für die Bewertung und Einordnung von Aussagen in positive, negative und neutrale Statements verantwortlich. Diese Form der Klassifikation wird das Kernelement der Börsenvorhersagen bilden.

Sentimentanalysen können entweder semantik- oder lernbasiert durchgeführt werden. Der Semantik Ansatz ordnet Wörter anhand eines Lexikons zu ihrer semantischen Ausprägung zu. Der lernbasierte Ansatz verwendet eine Reihe von manuell zugeordneten Aussagen als Trainingsdatensätze und vergleicht diese Hilfe von Ähnlichkeitsmodellen zu noch unklaren Aussagen. Für die genaue Durchführung dieser Analysen gibt es jedoch eine Vielzahl verschiedener Möglichkeiten beziehungsweise Ausprägungen. Es soll hier lediglich einige mögliche Vorgehensweise beschrieben werden. Der Grad an Meinungsausprägung, der auch als Polarität bezeichnet wird, kann entweder als eindeutige Zuordnung geschehen oder mit einer gewissen Skalierung. Als Sentimentträger kommen Adjektive, Verben und Adverbien in Frage.

4.1 Vorverarbeitung

Der erste Schritt einer Sentimentanalyse stellt die Aufbereitung des Textmaterials dar. Je nach Datenquelle können hier unterschiedliche Schritte anfallen. Lei Zhang u.a. schlagen für Twitter folgende Vorverarbeitungsschritte vor vgl. (Zhang, et al., 2011).

1: Das entfernen von Retweets

2: Abkürzungen in ihre ungekürzte Form überführen (z.B. wknd zu weekend)

3: Links und Benutzernamen aus den Tweets löschen.

Gary Beverungen und Jugal Kalita schlagen weitere Vorverarbeitungsschritte vor vgl. (Beverungen & Kalita, 2011)

4: Das Löschen aller Tweets, die nicht in Englisch verfasst sind.

5: Das Löschen von SPAM Tweets

Desweiteren empfehlen einige wissenschaftliche Quellen Verneinungen wie das englische „not“ im Text zu entfernen und es dafür direkt an das folgende Wort anzuhängen. Aus „not funny“ wird somit „not_funny“. Hierdurch soll vermieden werden, dass das positive Wort „funny“ nicht fälschlicherweise als positiv gewertet wird. Man vereinfacht hierdurch die Analyse, da man bei Verneinungen nur einzelne Wörter betrachten muss und nicht ihre Verbindung zueinander.

Zhang u.a. geben weiterhin an, dass in Tweets folgende drei Satzformen enthalten sein können.

Deklarative Sätze: Diese geben den Standpunkt des Autor wieder (z.B. Dies ist ein gutes Telefon).

Imperative Sätze: Diese geben einen Hinweise oder einen Befehl (z.B. Kauft nicht dieses Telefon).

Interrogative Sätze: Diese stellen eine Frage (z.B. Welches ist das beste Telefon).

Sowohl deklarative als auch imperative Sätze beinhalten sehr oft Meinungsäußerungen. Interrogative dagegen fragen nach einer Meinung, beinhalten selbst jedoch meisten keine. Es stellt sich also die Frage, ob diese Sätze gelöscht werden sollten.

Im Textmining wird bei der Vorverarbeitung für gewöhnlich eine Satzsegmentierung durchgeführt, bei der alle Sätze aus einem Text voneinander getrennt werden. Diese macht bei einer Untersuchung von Tweets jedoch recht wenig Sinn, da diese meist nur aus einem Satz bestehen.

Nun liegt das Rohmaterial für die weiteren Analysemethoden vor. Hierzu werden Methoden des Natural Language Processing (NLP) genutzt. Twitternachrichten haben jedoch gewisse Eigenheiten, die den Einsatz von NLP Methoden erheblich erschweren. Viele dieser Techniken wurden für längere Texte entwickelt, Twitternachrichten bestehen jedoch wie bereits erwähnt aus höchstens 140 Zeichen. Desweiteren beinhalten sehr viele Tweets Umgangssprache und grammatikalisch falsche Sätze. Außerdem werden Wörter entweder absichtlich oder unabsichtlich falsch geschrieben. Hinzu kommen gewisse Internetslangbegriffe wie lol, wtf oder fail, die sehr großen Einfluss auf die Meinungshaltung haben können. Schlussendlich hat Twitter auch noch seine eigenen Begriffe wie Benutzernamen (@Username), Retweets (RT) oder Trending Topics (TT). Die

Berücksichtigung all dieser Faktoren ist kein leichtes Unterfangen. Es muss daher abgewägt werden, ob man derartige Dinge berücksichtigt.

Das nun vorliegende Textmaterial wird in seine einzelnen Wörter zerlegt, um diese einzeln betrachten zu können. Dieser Schritt wird als Tokenisierung bezeichnet.

Darauf folgt, dass Part of Speach Tagging. Dieses hat die Aufgabe die Wortformen des vorliegenden Textes zu bestimmen. Es wird also bestimmt ob es sich bei einem Wort um ein Nomen, ein Verb, ein Adjektiv, Pronomen, Präpositionen usw. handelt. Dies ist notwendig da je nach Wortart ein Begriff eine andere Bedeutung haben kann. Es erfolgt also eine getrennte Betrachtung eines Wortes je nach Wortart.

Die nun vorliegenden Wortarten werden dann in ihre Stammform versetzt. Dieser Schritt wird als Stemming bezeichnet. Hierdurch verringert sich die Menge an möglichen Wörtern, da nur eine Konjugationsform berücksichtigt wird. Die bisher vorgestellten Methoden sind für die semantik- und lernbasierten Ansätze nahezu identisch. Bei den darauf folgenden Schritten unterscheiden sie sich jedoch. Als erstes soll nun der semantische Ansatz vorgestellt werden.

4.2 Semantikbasierter Ansatz

Dieser Ansatz vergleicht die im Text vorkommenden Worte mit einem Lexikon in dem alle meinungstragenden Worte samt deren Stimmungsrichtung verzeichnet sind. Diese Vorgehensweise wir auch als Bag of Words bezeichnet, da es die Reihenfolge der im Text vorkommenden Wörter nicht berücksichtigt, sondern diese lediglich als eine ungeordnete Menge betrachtet.

Um ein derartiges Opinion Lexikon zu erstellen, gibt es verschiedene Ansätze. Ein wichtiger Aspekt bei der Erstellung eines Opinion Lexikons ist die Berücksichtigung des Anwendungskontextes. Je nachdem über wen oder was man eine Aussage trifft, können andere Wörter zum Einsatz kommen. Es macht beispielsweise einen Unterschied, ob man über einen Menschen, eine Partei, eine Firma oder ein Produkt spricht. Beispielsweise kann der Begriff „gering“ in unterschiedlichen Sätzen eine positive oder negative Auswirkung haben. Bei einem Auto wäre ein geringer Verbrauch sehr positiv während eine geringe Leistung eher negativ ist. Eine Möglichkeit ist entsprechende Wörter manuell zu suchen und einzugeben. Bei der riesigen Anzahl an möglichen Wörtern stellt dies jedoch einen enormen Zeit und Arbeitsaufwand dar. Desweiteren sind manuell erstellte Sentiment-Lexika bis zu einem gewissen Grad subjektiv, da jeder Mensch eine etwas andere Auffassung bei der Zuordnung von Wörtern hat.

Eine Möglichkeit ein Opinion Lexikon semi-automatisch zu erstellen, ist eine Gruppe von eindeutig positiv oder negativen Wörter zu definieren und alle Synonyme dieser Wörter in die gleiche Gruppe einzuordnen. Diese manuell vorgegebenen Wörter werden auch als Samen (eng. Seeds) bezeichnet, da diese von einem kleinen Keim auf eine große Menge an Wörtern anwachsen. Die Wahl dieser Wörter muss gut überlegt sein, zumal die Qualität des hierdurch erzeugten Opinion-Lexikons von ihnen abhängt. Um eine möglichst gute Qualität zu erreichen, sollten Wörter oder Zeichen verwendet werden, die kontextunabhängig und absolut eindeutig sind. Für die Ausweitung dieser Samenkörner wird nun in einem Lexikon nach Synonymen der Wörter gesucht. Der Nachteil dieser Methode besteht darin, dass die

Ähnlichkeit eines Wortes mit der Distanz zum ursprünglichen Samenkorn abnimmt. Namrata Godbole u.a. gehen davon aus, dass die Zugehörigkeit eines Wortes mit der Tiefe des Pfades zum Samenkorn exponentiell abnimmt vgl. (Godbole, et al., 2007). Desweiteren geben die an, dass der letztendliche Wert die Summe aller Werte über alle Pfade darstellt. Pfade, bei denen ein Wort zwischen positiv und negativ wechselt, werden als unsicher angesehen. Für eine finale Aussage sollten für die Bewertung eines Wortes daher möglichst viele Pfade, die über dieses Wort führen, mit einbezogen werden. Je weniger ein Wort die Meinungsrichtung wechselt, desto vertrauenswürdiger ist es. Godbole u.a. konnten in ihrer Untersuchung 18000 meinungstragende Wörter durch eine kleine Gruppe an Samenkörnern bestimmen. Die berechneten Score-Werte der einzelnen Wörter folgten hierbei einer Normalverteilung. Die meisten Wörter lagen in der Mitte der Verteilung und konnten somit nicht eindeutig einer Meinungsrichtung zugeordnet werden. Derartige Wörter sollten daher für ein Opinion-Lexikon nicht verwendet werden. Besonders vielversprechend sind Wörter, die sich an den Extrempunkten der Normalverteilung befinden, da diese mit sehr hoher Wahrscheinlichkeit zu einer Meinungsrichtung und mit einer extrem niedrigen Wahrscheinlichkeit zu der anderen gehören.

Eine derartige Vorgehensweise lässt jedoch die Stärke der einzelnen Pfade außer Acht. Eine Möglichkeit dies zu ändern, ist die Einführung von Gewichtungen. Um diese zu erzeugen, bietet sich eine Vorgehensweise an, die Googles Page Rank Algorithmus ähnelt. Dieser nutzt die Linkstruktur des Webs um ein qualitatives Ranking von Webseiten zu erstellen.

Die meisten wissenschaftlichen Quellen verwenden für die Erstellung eines Opinion-Lexikons für englische Texte das so genannte WordNet[13] der Princeton Universität. WordNet ist eine große Datenbank der englischen Sprache. Sie gruppiert Wörter nach deren Wortform wie Nomen, Verben und Adjektiven. In jeder dieser Gruppen sind Wörter mit ähnlicher Bedeutung, konzeptioneller Semantik, semantischer und lexikalischer Verbindung verlinkt. Eine für Sentiment-Analysen aufbereitete Version des WordNet ist das SentiWordNet. In diesem Lexikon wurden allen Wörtern bereits ein Grad an Positivität, Negativität und Neutralität zugewiesen. Die Zuordnung der einzelnen Wörter wurde mit Hilfe der Verlinkungen der einzelnen Wörter erreicht. Man definierte ein Wort manuell als positiv, negativ oder neutral und ordnete alle ihm verwandten Wörter und Synonyme der gleichen Ausprägung zu. Jeder dieser Eigenschaften wird mit einer numerischen Skala zwischen null und eins bewertet. Die Summe der einzelnen Werte darf jedoch zusammen nicht mehr als eins ergeben. Pos. Score(term) + Neg. Score(term) + Objective Score(term) = 1.

Jedes positive und negative Wort eines Dokumentes wird nun gezählt und aufsummiert, um daraus die Meinungsrichtung des gesamten Dokumentes zu berechnen. Bruno Ohana und Brendan Tierneyy konnten unter Verwendung des SentiWordNet eine Klassifikationsgenauigkeit von 65,85% erreichen vgl. (Ohana & Tierneyy, 2009). Sie weisen darauf hin, dass ihre Methode damit ähnliche Ergebnisse erzielt hat wie manuell erstellte Sentiment-Lexika, deren Genauigkeit sie mit 69% angeben.

Das Zählen von Wörtern und deren Meinungsrichtung ist die einfachste Form des semantikbasierten Ansatzes. Neben diesem Ansatz existieren jedoch noch andere

[13] http://wordnet.princeton.edu/

artverwandte Möglichkeiten, die im Folgenden vorgestellt werden sollen. Casey Whitelaw schlagen die Verwendung so genannter Appraisal Groups vor vgl. (Whitelaw, et al., 2005). Darunter verstehen sie: „An appraisal group is represented as a set of attribute values in several task-independent semantic taxonomies, based on Appraisal Theory.“ Also eine Gruppe von Wörtern, die eine Meinung ausdrückt wie beispielsweise „nicht sehr gut“ oder „absolut hervorragend“. Die Appraisal Theorie, zu Deutsch Emotionstheorie, versucht zu erklären, was Emotionen sind, wodurch sie verursacht werden und wie sie sich auf das Verhalten von Menschen auswirken. Neben den Appraisal Groups wurden in deren Arbeit auch Appraisal Modifier berücksichtigt. Dies sind Wörter, die den Sinn einer Appraisal Group verändern wie „nicht“, „sehr“ oder „weniger“. Es stellte sich bei dieser Untersuchung jedoch heraus, dass die Berücksichtigung dieser Modifizierer die Klassifikationsgenauigkeit nicht wesentlich verbessern konnte. Die Autoren geben an, dass dies an der geringen Menge an meinungsändernden Wörtern in ihrer Untersuchung lag, die mit 1,8% extrem niedrig ausfiel.

Xiaowen Ding u.a. schlagen für die Erstellung eines universalen Sentiment-Lexikons einen ganzheitlichen Ansatz vor vgl. (Ding, et al., 2008). Ziel ihrer Entwicklung war es, kontextabhängige Meinungswörter richtig zuzuordnen. Als Grundlage ihrer Entwicklung dient das so genannte „feature based“ Opinion-Mining. Diese Analyseform entspricht dem bereits vorgestellten Opinion-Mining- Modell, welches versucht Meinungsäußerungen den Features von Objekten zuzuordnen. Es wird also nicht mehr nur ein Sentiment Wort betrachtet, sondern der Zusammenhang zwischen diesem Wort und einem Feature.

Um eine Meinungsrichtung zu bestimmen, werden die Sentiment-Wörter, die nahe an einem Feature stehen, für jeden Satz gezählt. Wenn die Anzahl an positiven Wörtern überwiegt, wird der Satz als positiv gewertet, ansonsten ist er negativ. Da es keine Möglichkeit gibt, nur aufgrund eines kontextabhängigen Sentiment-Wortes und des Features, das es beeinflusst, deren Meinungsausrichtung zu bestimmen, müssen vor der Bewertung Kenntnisse über das Objekt und deren Features bekannt sein. Die Autoren schlagen hierzu die Einbeziehung von externen Informationen, anderen Sätzen, anderen Texten sowie gewissen Spracheigenschaften vor.

Es wurde hier desweiteren eine Unterscheidung zwischen expliziten und impliziten Meinungsäußerungen getroffen. „An explicit opinion on feature f is a subjective sentence that directly expresses a positive or negative opinion. An implicit opinion on feature f is an objective sentence that implies an opinion” (Ding, et al., 2008). Ein Beispiel für einen expliziten Satz wäre: Die Fahreigenschaften des Autos finde ich hervorragend. Ein Beispiel für einen impliziten Satz könnte sein: Das Auto war schon nach zwei Wochen kaputt.

Die Autoren verwenden nun ein Modell, in dem jedes Objekt mit einer endlichen Menge an Features beschrieben werden kann. Jedes Feature kann nun mit einer bestimmten Menge an Wörtern bewertet werden. Zu jedem Wort gibt es desweiteren eine endliche Menge an Synonymen, die eine gleichwertige Meinung ausdrücken.

Um eine entsprechende Analyse durchführen zu können, müssen das Objekt, dessen Features sowie die meinungstragenden Wörter in dem zu untersuchenden Text bekannt sein. Das Ergebnis einer auf diesem Modell basierende Analyse, ist ein Reihe von Feature- Meinungs-

Paaren, bei denen jedes Feature eine Meinungsrichtung durch die ihm zugeordneten Wörter bekommt. Eine Möglichkeit das Modell zu verbessern, wäre das Hinzufügen von Gewichtungen für die meinungstragenden Wörter. Diese Methode wurde in der Untersuchung von Ding u.a. jedoch nicht berücksichtigt.

Um den Kontext eines Meinungstragenden Wortes festzustellen, wenden die Wissenschaftler folgende Vorgehensweise an. Sie versuchen bei Wörtern bei denen nicht klar ist, ob diese positiv oder negativ gemeint sind, andere Sätze zum gleichen Objekt zu finden, bei denen das Wort in einem Zusammenhang mit einem in eine eindeutige Richtung weisenden Meinungswort steht. Beispielsweise ein Wort wie „lange“ in einem Satz wie: Die Kamera macht großartige Fotos und hat eine lange Akkulaufzeit. Da lange in einem Satz mit dem Wort großartig steht, wird davon ausgegangen, dass es ebenfalls positiv gemeint ist. Es wird dabei die Vermutung zu Grunde gelegt, dass ein Satz eine einzige Meinung vertritt, so lange kein meinungsumkehrendes Wort wie beispielsweise das Englische „but“ vorkommt, dass alle folgenden Äußerungen in ihrer Meinung umkehrt. Ein Beispiel Satz hierfür wäre: „The Camera makes great pictures but has a short battery life.“ Ein weiteres Merkmal, das man sich hier zunutze macht, ist die Tatsache, dass Menschen positive und negative Äußerungen meist zusammen gruppieren. Der Anfang von Satzgruppen, die eine andere Meinung vertreten, wird meist von Wörtern wie „allerdings“, „leider“ oder „jedoch“ eingeleitet. Wenn ein Wort in einem speziellen Kontext nun als positiv oder negativ identifiziert wurde, wird seinen Synonymen die gleiche Meinungsrichtung zugeordnet. Seinen Antonymen wird hingegen die entgegengesetzte Meinungsrichtung zugeordnet.

Für die Gesamtbewertung eines Dokumentes wurden in dieser Untersuchung nun allen positiven Wörtern ein Wert von +1 sowie allen negativen Wörtern ein Wert von -1 zugeordnet. Alle Werte wurden schließlich mit folgender Formel zusammengefasst.

$$score(f) = \sum_{w_i : w_i \in s \wedge w_i \in V} \frac{w_i.SO}{dis(w_i, f)},$$

Hierbei ist wi ein Opinion-Wort, V ist die Menge aller Opinion-Wörter und s der Satz, der die Features enthält. dis(wi, f) steht für die Distanz zwischen dem Opinion-Wort und dem Feature im jeweiligen Satz. wi.SO ist die semantische Orientierung des Wortes wi. Die Distanzfunktion im Nenner wird dazu verwendet um Wörtern, die eine hohe Distanz zu einem Feature aufweisen, eine geringere Gewichtung zuzuweisen. Dieser Vorgehensweise liegt die Idee zu Grunde, dass Wörter die weit weg von einem Feature stehen, womöglich sich nicht auf dieses beziehen. Wenn der finale score Wert positiv ist, wird das Feature f im Satz s als positiv angesehen. Bei einem negativen Wert wird es entsprechend als negativ bewertet. Die Autoren weisen jedoch darauf hin, dass diese Aggregationsform nicht mit Vergleichsätzen wie beispielsweise „Audi ist besser als BMW“ funktioniert. Derartige Sätze weisen einem Objekt eine positive Orientierung zu, während einer anderen eine negative zugewiesen wird.

4.3 Lernbasierter Ansatz

Neben dem Semantikansatz existiert wie bereits erwähnt, der lernbasierte Ansatz. Dieser nutzt Textklassifikationsmethoden des Information Retrieval, Klassifikationsverfahren des Data Mining und kann dem maschinellen Lernen zugeordnet werden. Maschinelles Lernen versucht eine Entscheidungsaufgabe aufgrund von gesammelter Erfahrung zu treffen. Ziel ist das Erlernen von Zusammenhängen zwischen Zielwerten und den Eigenschaften von Objekten. Slobodan Vucetic beschreibt das Feld des maschinellen Lernens wie folgt:

„The field of machine learning studies the design of computer programs able to induce patterns, regularities, or rules from past experiences. Learner (a computer program) processes data D representing past experiences and tries to either develop an appropriate response to future data, or describe in some meaningful way the data seen" (Vucetic, 2003).

Ziel ist eine automatische Klassifikation der Daten in gewissen Zielgruppen vorzunehmen. Die Klassifikation versucht dabei ein Objekt aufgrund seiner Merkmale in bekannte Klassen einzuordnen. Die zu untersuchenden Objekte besitzen dabei eine Menge an Merkmalen, die diese beschreiben. Diese können entweder quantitativer oder qualitativer Natur sein. Ein quantitatives Merkmal zeichnet sich dadurch aus, dass es prinzipiell jeden Wert in einem endlichen oder unendlichen Intervall annehmen kann. Beispiele für derartige Merkmale sind beispielsweise Körpergröße, Lebensalter oder Temperatur. Im Gegensatz dazu, besitzt ein qualitatives Merkmal eine endliche Zahl von Alternativen oder Zuständen. Beispiele hierfür können Farben, Monatsangaben oder das Geschlecht sein. Es ist üblich, derartige Merkmale mit Zahlen zu kodieren (rot =1, blau=2, gelb =3 usw.). Derartige Merkmale können, müssen aber keine Ordnung besitzen. Qualitative Merkmale, die nur zwei Ausprägungen besitzen, werden als binär bezeichnet.

Die gebräuchlichste Methode um eine Klassifikation von Texten zu beginnen, ist eine Datenmatrix zu erstellen, in der alle Objekte samt deren Merkmalen und Ausprägungen eingetragen werden. Folgende Matrix zeigt die Eigenschaften von zwei Dokumenten. Die Eigenschaften geben dabei an, ob dieser Begriff im Dokument vorhanden ist (1) oder nicht (0). In diesem Beispiel ob gewisse Freizeitaktivitäten erwähnt werden.

Datensatz	Segeln	Inline-Skaten	Surfen	Aerobic	Jogging
1.	1	0	0	0	1
2.	1	1	1	0	1

Ausgehend von dieser Matrix wird nun versucht, Ähnlichkeiten der einzelnen Objekte zu berechnen. Neben der Ähnlichkeit der Objekte lässt sich auch deren Unähnlichkeit (Distanz) messen. Beide Maßeinheiten sind dabei aufeinander abbildbar. Es ist dabei möglich, wenn auch nicht unbedingt notwendig, die Objekte auf ein einheitliches Maß zu skalieren. In den meisten Fällen wird dabei ein Übereinstimmungsmaß zwischen 0% und 100% gewählt. Die Zielklassen können hierbei drei Formen haben. Disjunkte Klassen, bei denen ein Objekt nur in einer Klasse existieren darf, nicht disjunkte Klassen, bei denen ein Objekt in mehreren Klassen existieren darf und hierarchischen Klassen bei denen eine Ordnung existiert.

Um die Ähnlichkeit der einzelnen Matrizeneinträge zu berechnen, gibt es verschiedene Arten, die auf empirischen Methoden beruhen oder aus der Statistik stammen. Diese Methoden werden als Ähnlichkeits- bzw. Distanzmaße bezeichnet. Sie dienen dazu, die Ähnlichkeit respektive Unähnlichkeit zweier Objekte zu quantifizieren. Beide Maße untersuchen, wie nahe die Eigenschaften zweier Objekte beieinanderliegen. Man geht davon aus, dass ähnliche Objekte den gleichen, und unähnliche Objekte verschiedenen Gruppen angehören. Ingo Schmitt definiert ein Ähnlichkeitsmaß wie folgt.

„Ein Ähnlichkeitsmaß ist eine Funktion, die einem Paar von Objekten eine Zahl aus dem reellen Intervall [0; 1] zuordnet. Dabei korrespondiert der Wert 1 zur maximalen Ähnlichkeit und der Wert 0 zur maximalen Unähnlichkeit“ (Schmitt, 2005).

Kein Ähnlichkeitsmaß kann jedoch durchweg als das beste angesehen werden. Das jeweilige Ähnlichkeits- bzw. Distanzmaß muss sich nach dem jeweiligen Fall richten und die Semantik der Daten berücksichtigen. Da die Sentiment-Analyse mit binären Ähnlichkeitsmodellen arbeiten wird, werden diese hier näher erläutert.

Binäre Variablen können nur zwei Zustände annehmen, null oder eins. Diese Zustände werden meist dafür verwendet, um anzugeben, ob eine Eigenschaft erfüllt ist oder nicht. Bei dem Vergleich zwischen zwei Objekten mit binären Eigenschaften, können nur vier verschiedene Fälle auftreten 1/1, 1/0, 0/1 oder 0/0. Letztendlich läuft ein derartiger Vergleich immer darauf hinaus, ob zwei Eigenschaften entweder gleich oder ungleich sind. Jegliche Maße die derartige Objekte miteinander vergleichen, nutzen die Häufigkeiten mit denen die einzelnen Wertekombinationen auftreten und vergleichen diese mit der Gesamtzahl der Wertepaare. Die verschiedenen Ähnlichkeitsmaße unterscheiden sich lediglich in der Art und Weise wie diese Häufigkeiten miteinander verglichen werden. Ähnlichkeitsmaße vergleichen Wertekombinationen mit gleichen Werten wie 1/1 und 0/0, Unähnlichkeitsmaße bewerten entsprechend die ungleichen Wertpaare 0/1 und 1/0.

4.3.1 Distanzmaße

Das am häufigsten verwendete Distanzmaße ist die euklidische Distanz. Der euklidische Abstand zweier Punkte in der Ebene oder im Raum entspricht der direkten Linie zwischen zwei Punkten in einem Vektorraum. Abbildung 8 zeigt die Vektoren A und B in einem zweidimensionalen Vektorraum. Die euklidische Distanz entspricht der Strecke d zwischen die Punkten A und B.

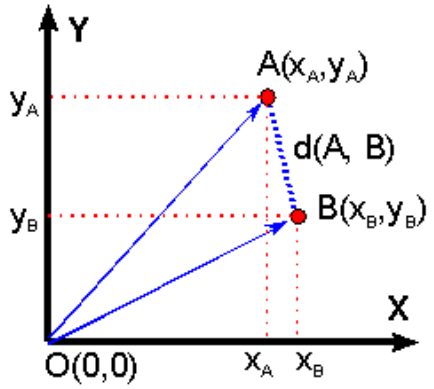

Abbildung 8: Euklidische Distanz[14]

Nach diesem Maß ergibt sich die Distanz zwischen den beiden Objekten X und Y nach der allgemeinen Formel.

$$\text{Distanz}_{X,Y} = \sqrt{\sum_{i=1}^{v} (X_i - Y_i)^2} \quad .$$

Es wird die Differenz zweier Dimensionen eines Objektes quadriert und mit der Quadrierten Differenz aller anderen Dimensionen addiert. Von diesem Wert wird schließlich die Wurzel genommen.

Die Folgende Matrix zeigt die vier möglichen Zustände bei dem Vergleich von binären Merkmalen.

		Objekt 1	
		Erfüllt	Nicht erfüllt
Objekt 2	Erfüllt	a	b
	Nicht erfüllt	c	d

Wenn man das Konzept der euklidischen Distanz auf binäre Merkmale überträgt, vereinfacht sich die Formel zur Berechnung der Distanz. Da die Differenz zweier gleicher Merkmale (1-1) immer null ist, fallen diese Werte weg. Somit müssen nur alle ungleichen Wertepaare

[14] Quelle: http://ifgivor.uni-muenster.de/vorlesungen/Geoinformatik/kap/kap4/img00016.gif

betrachten werden. Die Differenz zweier gleicher binärer Merkmale (0-1, 1-0) ist immer eins. Da die Quadrierung von eins wiederum eins ergibt, ist die Quadrierung überflüssig. Die euklidische Distanz binärer Merkmale ist somit gleich der Quadratwurzel aus der Summe ungleicher Wertepaare:

$$\text{Euklidische Distanz} = \sqrt{b + c}$$

4.3.2 Klassifikationsalgorithmen

Die Klassifikation versucht eine Gruppeneinteilung von Datensätzen aufgrund deren Eigenschaften zu treffen. In diesem Fall ob ein Dokument positiv oder negativ ist, basierend auf den Fragmenten des Dokuments. Die Güte eines derartigen Klassifikators wird vor allem dadurch gemessen, inwiefern geplante und beobachtete Gruppenzugehörigkeit übereinstimmen. Für die Bestimmung der Gruppenzugehörigkeit gibt es verschiedene Verfahren.

Am Anfang einer derartigen Klassifikation steht die Erstellung von Trainingsdaten. Dies sind Datensätze bei denen die Entscheidung für die Wahl der Zielvariablen bereits getroffen wurde. In diesem Fall Dokumente, die bereits in positiv, negativ oder neutral eingeordnet wurden. Derartige Meinungszuordnungen werden meist manuell von Menschen durchgeführt. Bei manuell erstellen Trainingsdaten spricht man von „supervised learning“. Diese Trainingsdaten dienen dem Klassifikationsalgorithmus danach als Entscheidungsgrundlage. Eine semiautomatische Methode, die einige wissenschaftliche Quellen hierfür anwenden, ist die Verwendung von Emoticons (z.B. :-), :-D, :-[ ) oder sentiment-bezogene Hashtags wie #fail, #success oder #epic. Da diese Zeichenfolgen ein Dokument relativ sicher als positiv oder negativ klassifizieren, eigenen sie sich sehr gut für die Auswahl von Trainingsdatensätzen.

Dabei stellt sich die Frage, welche Daten in die Entscheidung mit eingebunden werden sollen, denn nicht alle Datensätze besitzen die gleiche Anzahl an Informationen. „Je feiner ein Klassifikator auf die Detailstruktur der Beispieldaten eingeht, desto weniger generalisierungsfähiger ist er“ (Uhr, et al., 2003). Diese Überspezialisierung bezeichnet man auch als Overfitting. Hierbei kann zwar die Zielvariable durch die Zuhilfenahme sehr vieler Variablen sehr gut vorausgesagt werden, allerdings trifft der daraus entstandene Algorithmus nur noch auf eine sehr kleine Zahl von Datensätzen zu.

Die Erstellung der Testdaten ist sehr kostspielig, was dazu führt, dass die Anzahl an Trainingssätzen im Verhältnis zur Anzahl der Testdatensätze meist relativ gering ausfällt. Die Aufgabe eines Trainingsalgorithmus besteht darin, eine Entscheidungsregel aus den Daten abzuleiten, mit der sich die einzelnen Datensätze in die Zielgruppen einteilen lassen. Diese Entscheidungsregel wird auch als Klassifikator bezeichnet. Eine wichtige Eigenschaft des Klassifikators ist die Generalisierungsfähigkeit. Diese bestimmt wie gut der Algorithmus mit ihm ungekannten Daten funktioniert. Ziel ist eine möglichst hohe Generalisierungsfähigkeit, da diese zu einer kleinen Klassifikationsfehlerrate führt. Um diese Eigenschaft zu überprüfen, muss der Klassifikator auf eine Reihe von Testdatensätzen angewendet werden. Es gibt

verschiedene mathematische Ansätze, um einen derartigen Klassifikator zu erstellen. Einige der wichtigsten werden im folgenden vorgestellt.

Support Vektor Maschinen (SVM)

Diese Art der Klassifikation versucht die Einteilung von Objekten in Klassen durch ihre Position in einem N-Dimensionalen Vektorraum zu bestimmen. Ausgangspunkt ist eine Menge von Trainingsdaten, deren Klassenzugehörigkeit bekannt ist. Jedes dieser Objekte wird als Vektor in einem Vektorraum repräsentiert. In diesem Vektorraum gilt es nun eine Linie oder Fläche zu finden, welche die einzelnen Klassen möglichst sauber voneinander trennt. Diese Trennfläche wird als Hyperebene bezeichnet. Ziel ist es eine Hyperebene zu finden, welche den Abstand zwischen ihr und den einzelnen Klassen möglichst maximiert. Das Problem einer Ebene ist allerdings, dass sie nur eine lineare Trennung vornehmen kann. Dies ist in den meisten Fällen jedoch nicht möglich. Um in derartigen Fällen trotzdem eine Trennung vornehmen zu können, werden die Objekte in einen höherdimensionalen Raum überführt (siehe Abb. 9). Die Anzahl an Dimensionen kann dabei theoretisch unendlich groß sein. Hierbei werden solange neue Dimensionen hinzugefügt, bis eine lineare Trennebene zwischen den einzelnen Klassen gefunden werden kann. Bei einer Rücktransformation dieser hochdimensionalen Hyperebene in eine niedrig dimensionale ergeben sich nicht lineare Trennflächen die gegebenenfalls auch nicht zusammenhängen. Diese sind nun in der Lage die einzelnen Klassen sauber voneinander trennen.

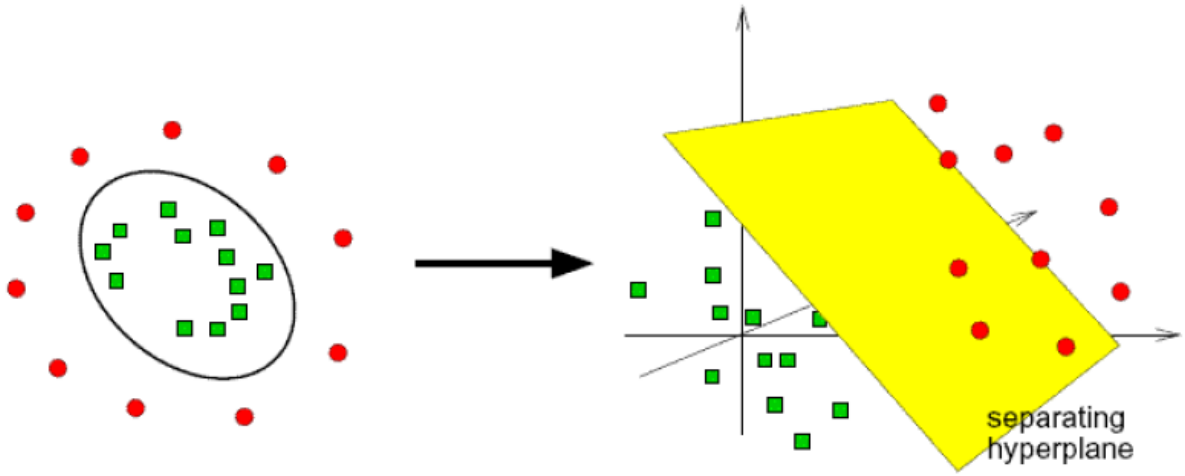

Abbildung 9: SVM Hochtransformierung Hyperebene[15]

Die einzelnen Dimensionen stellen dabei die Eigenschaften eines Objektes dar. In einer Textklassifikation sind die Eigenschaften eines Dokumentes die Wörter die es enthält. Jedes Wort ist also eine eigenständige Dimension. Da hier wiederum nur das Vorhandensein eines Wortes gewertet werden kann, liegen lediglich binäre Dimensionen vor.

[15] Quelle: http://cvpr.uni-muenster.de/teaching/ws06/mustererkennungWS06/script/ME11.pdf

K-Nearest Neighbor (KNN)

KNN ist ein instanzenbasiertes Lernverfahren, das eine Zuordnung von Objekten zu einer Klasse durch ihre Nähe zu anderen Objekten herstellt, deren Klassenzugehörigkeit bereits bekannt ist. Ein Objekt wir also anhand der Nähe zu seinen nächsten Nachbarn klassifiziert (siehe Abb. 10). Wie bei den Support Vektor Maschinen werden auch hier die Objekte als Vektoren in einem hochdimensionalen Vektorraum betrachtet. Für die Zuordnung der zu klassifizierenden Objekte zu anderen Objekten wird die Distanz zwischen ihnen gemessen. Hierfür wird meist die euklidische Distanz verwendet.

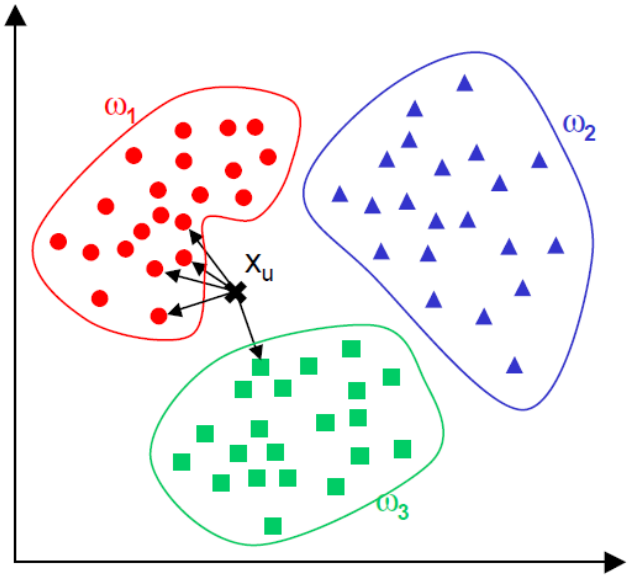

Abbildung 10: KNN Klassifikation[16]

Naive Bayes

Die Naive Bayes Klassifikation versucht durch die Berechnung bedingter Wahrscheinlichkeiten eine Zuordnung von Datensätzen vorzunehmen. Hierfür wurde vom englischen Mathematiker Thomas Bayes eine spezielle Methode entwickelt. Der Begriff Naiv leitet sich aus der Tatsache heraus, dass bei dieser Methode angenommen wird, dass alle Attribute eines Datensatzes voneinander unabhängig sind. Diese Annahme ist zwar unrealistisch, jedoch konnte die Naive Bayes Methode in der Praxis sehr gute Ergebnisse erzielen. Letztendlich liefert ein Bayes Klassifikator eine Wahrscheinlichkeit der Zugehörigkeit eines Datensatzes zu einer Gruppe. Die Wahrscheinlichkeit berechnet sich dabei wie folgt.

[16] Quelle: http://courses.cs.tamu.edu/rgutier/cs790_w02/l8.pdf

Wir betrachten eine Menge von Trainingsdaten mit d Datensätzen, die jeweils n Attribute besitzen, eingeteilt in k Klassen. Für jedes n-te Attribut wird nun eine Wahrscheinlichkeit durch die Häufigkeit seines Auftretens im Verhältnis zum Auftreten einer Klasse in der Gesamtmenge der Dokumente berechnet. Dies geschieht mit folgender Formel.

L[n] = Menge an Dokumenten in denen n und k vorkommen / Menge aller Dokumente in denen k vorkommt.

Dieser Wahrscheinlichkeitswert wird nun für alle Attribute berechnet. Die finale Wahrscheinlichkeit der Zugehörigkeit eines Dokumentes zu einer Klasse berechnet sich schließlich aus dem Produkt aller Attribut Wahrscheinlichkeiten, multipliziert mit der Häufigkeit des Vorkommens einer Klasse in allen Dokumenten. Folgendes Beispiel, das einem Data Mining Glossar der Universität Kaiserslautern entnommen wurde, soll dies noch einmal verdeutlichen siehe (AG Algorithmisches Lernen, 2001).

Gegeben sei eine Reihe von Datensätzen, welche die Wohnsituation einiger Menschen beschreibt. Für jede Person ist bekannt, ob diese ein Arbeitsverhältnis als Angestellter hat oder selbstständig ist. Desweiteren ist bekannt, ob jene Person verheiratet oder ledig ist und ob diese Kinder besitzt. Für die hier vorliegenden Trainingsdatensätze ist ebenfalls bekannt, ob die Person Mieter oder Eigentümer der Wohnung ist. Ziel ist es aufgrund der Eigenschaften Beruf, Familienstand und Kinder vorherzusagen, ob diese Person Mieter oder Eigentümer der Wohnung ist.

Beruf	Familienstand	Kinder	Wohnung
angestellt	verheiratet	ja	Miete
angestellt	ledig	nein	Eigentum
angestellt	verheiratet	nein	Miete
angestellt	verheiratet	ja	Eigentum
selbständig	ledig	ja	Eigentum
selbständig	ledig	nein	Miete
selbständig	verheiratet	ja	Eigentum
selbständig	verheiratet	nein	Eigentum

Mit dieser Tabelle lassen sich folgende relativen Häufigkeiten berechnen. Die relative Häufigkeit von „Miete“ liegt beispielsweise bei 3/8, da drei von acht Personen zur Miete wohnen. Die bedingte Wahrscheinlichkeit, dass man als Angestellter ein Mieter ist, liegt bei 2/3, da von den drei Mietwohnungen zwei von angestellten bewohnt werden. Mit dieser Vorgehensweise lassen sich nun die bedingten Wahrscheinlichkeiten für alle Eigenschaften im Bezug auf das Wohnverhältnis erstellen.

Beruf	Familienstand	Kinder
a/M = 2/3	v/M = 2/3	j/M = 1/3
s/M = 1/3	l/M = 1/3	n/M = 2/3
a/E = 2/5	v/E = 3/5	j/E = 3/5
s/E = 3/5	l/E = 2/5	n/E = 2/5

Mit dieser Tabelle lassen sich nun die Wahrscheinlichkeiten der einzelnen Eigenschaften verknüpfen. Wenn nun berechnet werden soll, ob ein lediger Angestellter zur Miete wohnt oder eine Eigentumswohnung besitzt ergäbe sich folgende Gleichung. Den Wahrscheinlichkeiten der einzelnen Eigenschaften wird die Wahrscheinlichkeit der Klasse vorangestellt.

*L[Miete] = 3/8 * 2/3 * 1/3 * 1/3 = 1/36 = 0.028*
*L[Eigen] = 5/8 * 2/5 * 2/5 * 3/5 = 1/25 = 0.04*.

Daraus liegen die Wahrscheinlichkeiten der Zugehörigkeit Miete und Eigentum bei:

P[Miete] = 0.028 / (0.028 + 0.04) = 0.41
P[Eigen] = 0.04 / (0.028 + 0.04) = 0.59

Die Wahrscheinlichkeit für eine Mietwohnung liegt damit etwas höher als die einer Eigentumswohnung.

Maximum Entropie

Der Begriff der Entropie wird in verschiedenen wissenschaftlichen Disziplinen für unterschiedliche Dinge verwendet und definiert. Vielen dieser Definitionen ist jedoch gemein, dass der Begriff Entropie das Maß an Unordnung in einer endlichen Menge an Objekten beschreibt. Aus dieser Unordnung resultiert ein Grad an Ungewissheit über die Vorhersagbarkeit von Zuständen dieser Menge. Hierdurch wird Entropie auch oft als Maß an Unsicherheit verwendet. Eine geringe Entropie sagt demnach einen hohen Grad an Sicherheit, eine hohe Entropie ein hohes Maß an Unsicherheit aus. In der Klassifikation wird die Entropie Definition von Claude E. Shannon zu Grunde gelegt. Dieser definierte Entropie als das Maß an Unsicherheit in einer gegebenen Wahrscheinlichkeitsverteilung. Shannon stellte hierfür folgende Formel auf.

$$H(X) \quad = \quad -\sum_x p(x) \log_2 p(x)$$

Ziel ist eine Wahrscheinlichkeitsverteilung zu finden, die die Entropie maximiert und konsistent mit allen bekannten Informationen ist.

„The probability distribution that uniquely represents or encodes our state of information is the one that maximises the uncertainty measure H(p) while remaining consistent with our information“ (Schulte im Walde, et al., 2004).

Eine Grundregel der maximalen Entropie besteht darin, nur die Informationen zu verwenden, die sicher sind bzw. beobachtet werden können. Es sollte also alles modelliert werden, was bekannt ist und keine Annahmen darüber getroffen werden, was unbekannt ist. Es wird schließlich das Modell gewählt, dass alle Fakten berücksichtigt, jedoch alle Wahrscheinlichkeiten möglichst gleichmäßig verteilt. Das am häufigsten verwendete Klassifikationsverfahren, die Entscheidungsbäume, verwendet das Entropiemaß beispielsweise als Trennkriterium für die Einteilung von Objekten in Gruppen.

Ein Beispiel:

Betrachten wir eine Münze, deren Zustand beim Werfen der Münze (Kopf oder Zahl) vorhergesagt werden soll. Wenn wir eine faire Münze betrachten, die mit jeweils 50% Wahrscheinlichkeit Kopf oder Zahl sein kann, liegt eine maximale Unsicherheit vor. Der Entropie-Wert dieser Verteilung läge bei eins. Hätten wir eine Münze die mit 100% Wahrscheinlichkeit immer einen Zustand annehmen würde, läge eine vollkommene Sicherheit vor. Der Entropie- Wert dieser Verteilung läge bei null. Die untere Grenze für eine Wahrscheinlichkeitsverteilung liegt immer bei Entropie = null für vollkommene Sicherheit. Die obere Grenze dagegen ergibt sich bei einer vollkommenen Gleichverteilung aller Zustände.

Sollten verschiedene Wahrscheinlichkeitsmodelle zu einem Zufallsexperiment wie das des Münzwurfes vorliegen, verlangt das Prinzip der maximalen Entropie, dass grundsätzlich jenes gewählt wird, das die größte Entropie und somit die größte Unsicherheit besitzt.

In einer Maximum Entropie Klassifikation hängt die Wahrscheinlichkeit, dass ein Dokument zu einer bestimmten Klasse gehört davon ab, ob es die Entropie des Klassifikationssystems maximiert. Um eine Einordnung in eine Gruppe vorzunehmen, benötigt ein Maximum-Entropie-Klassifizierer eine Menge von Features, die diese Gruppe beschreiben. Im Falle der Sentimentanalyse sind dies meist die einzelnen Wörter eines Textes. Die Zugehörigkeit eines Wortes in einem Text zu einer Gruppe wird mit einer Wahrscheinlichkeit angegeben. Am Anfang einer derartigen Analyse liegt jedoch noch keine Wahrscheinlichkeitsverteilung vor. Diese wird durch die Trainingsdaten erstellt. Ziel der Trainingsdaten ist es eine Reihe von Wahrscheinlichkeitsbedingungen für jedes einzelne Wort zu erstellen.

Ein Beispiel für eine Textklassifikation mittels Maximum-Entropie entnommen aus (Berger, et al., 1996).

Nehmen wir an wir wollten das englische Wort „in“ ins französische übersetzen. Da es mehrere Wörter im französischen gibt, die diesem Wort entsprechen, gilt es gewisse Entscheidungsregeln zu finden, welches dieser Wörter zu wählen ist. Um diese Entscheidungsregeln zu erlangen, könnten wir einen Übersetzungsexperten zu Rate ziehen. Wir würden feststellen, dass dieser für die Übersetzung ausschließlich die Wörter „dans“, „en“, „á“, „au cours de“ und „pendant“ verwendet. Hieraus können wir die erste Entscheidungsregel ableiten, nämlich das die Wahrscheinlichkeit, dass das zu wählende Wort eines dieser Fünf sein muss bei 100% liegt.

$$p(\text{dans}) + p(\text{en}) + p(\text{á}) + p(\text{au cours de}) + p(\text{pendant}) = 1$$

Bis jetzt wissen wir allerdings noch nichts darüber, mit welcher Wahrscheinlichkeit die einzelnen Wörter gewählt werden. Da wir keine Annahme darüber treffen dürfen, was wir nicht wissen, müssen wir davon ausgehen, dass die Wahrscheinlichkeiten gleichmäßig verteilt sind und somit jede Wortform mit der Wahrscheinlichkeit von einem Fünftel vorkommt. Diese Annahme sorgt für die größtmögliche Unsicherheit und somit zur höchsten Entropie. Es ist jedoch sehr unwahrscheinlich, dass diese Annahme stimmt. Um die Wahrscheinlichkeitsverteilung entsprechend zu verbessern, gilt es neue Entscheidungsregeln

aufgrund der Wahl des Experten zu finden. Als nächstes würden wir feststellen, dass unser Experte in 30% der Fälle entweder „dans“ oder „en“ verwendet. Mit diesem Wissen könnten wir unsere Wahrscheinlichkeitsverteilung anpassen.

$$p(\text{dans}) + p(\text{en}) = 3/10$$

$$p(\text{dans}) + p(\text{en}) + p(\text{á}) + p(\text{au cours de}) + p(\text{pendant}) = 1$$

Da wir wiederum nicht wissen wie sich die Wahrscheinlichkeiten zwischen „dans“ und „en“ aufteilen, müssen wir wieder davon ausgehen, dass diese gleichmäßig verteilt sind. Somit nehmen wir an, dass jedes der beiden Wörter mit einer Wahrscheinlichkeit von 15% gewählt wird. Die restlichen 70% werden wiederum gleichmäßig auf die übrigen Wörter verteilt. Hieraus ergibt sich folgende Verteilung.

$p(\text{dans}) = 3/20$

$p(\text{en}) = 3/20$

$p(\text{á}) = 7/30$

$p(\text{au cours de}) = 7/30$

$p(\text{pendant}) = 7/30$

Diese Vorgehensweise wird nun soweit fortgesetzt bis die Wahrscheinlichkeitsverteilung konvergiert. Sicherlich ist die ausschließliche Betrachtung des Wortes „in“ ohne den Kontext, in dem es sich befindet, wenig sinnvoll, allerdings ist die Aufgabe einer Maximum Entropie Klassifikation ein stochastisches Modell zu finden, welches das Verhalten eines Zufallsexperimentes hinreichend genau beschreibt. In einer Sentiment-Analyse besagt die Wahrscheinlichkeitsverteilung welches Wort ein Dokument mit welcher Wahrscheinlichkeit als positiv oder negativ klassifiziert.

4.3.3 Anwendungsbeispiel Textklassifikation

Kommen wir nun zu einem konkreten Beispiel für die Anwendung einer Textklassifikation. Die ersten Schritte sind wie bereits beschrieben ähnlich dem semantikbasierten Ansatz. Der Text wird aufbereitet, seine Wortarten werden identifiziert und die einzelnen Wörter in ihre Grundform versetzt.

Als kleinste Unterscheidungseinheit verwendet dieser Ansatz sogenannte N-Gramm Modelle. Bei N-Grammen wird ein Text in Fragmente zerlegt. Fragmente können unter anderem Buchstaben oder Wörter sein. Die Anzahl an Wörtern oder Buchstaben, die zu einem Fragment zusammengefasst werden, kann dabei beliebig variiert werden. Bei der Bezeichnung der Fragmentgröße werden in der Regel griechische Zahlwörter verwendet. So werden Fragmente mit einem Zeichen oder Wort als Monogramme bezeichnet, Fragmente mit zwei Wörtern oder Zeichen entsprechend als Digramme usw. Wird eine Vielzahl von Fragmenten genutzt, spricht man meist von Multigrammen oder eben N-Grammen. Die hierzu genutzten Methoden werden als Tokenisierung bezeichnet, da sie den Text in einzelne Tokens aufspalten. Folgendes Beispiel stellt eine Aufspaltung eines Satzes in Monogramme dar.

Bsp. Gestern war ein kalter Tag -> Gestern + war + ein + kalter + Tag

Aus den hierdurch erstellten Fragmenten wird eine Term-Document-Matrix erstellt. Diese stellt alle Fragmente in einer Dimension und alle Dokumente in einer zweiten Dimension dar. Anders ausgedrückt haben wir eine Tabelle, in der in jeder Spalte ein Fragment steht und in jeder Zeile ein zu analysierendes Dokument. Dabei wird eine binäre Zuordnung getroffen. Wenn ein Fragment in einem Dokument vorhanden ist, wird der entsprechende Eintrag auf eins gesetzt ansonsten ist dieser null. Folgendes Beispiel soll dies noch einmal veranschaulichen.

Dokument A: Guten Tag

Dokument B: Gestern war ein kalter Tag

Dokument C: Schöner Tag heute

	Guten	Tag	Gestern	war	ein	kalter	Schöner	heute
Dokument A	1	1	0	0	0	0	0	0
Dokument B	0	1	1	1	1	1	0	0
Dokument C	0	1	0	0	0	0	1	1

Diese Art der Zuordnung beschreibt jedoch lediglich das Vorhandensein bestimmter Fragment in einem Dokument, es gibt jedoch keine Auskunft darüber, wie aussagekräftig ein bestimmtes Fragment eines Dokumentes gegenüber einem Suchbegriff ist. Eine Möglichkeit dies zu ändern, ist die Einführung von Gewichtungen gegenüber den einzelnen Fragmenten. Für eine derartige Gewichtung vorzunehmen, gibt es verschiedene Algorithmen. Eine sehr häufig verwendete Methode ist die Term Frequency - Inverse Document Frequency (TF-IDF). Die Term Frequenz beschreibt, wie häufig ein Wort in einem Dokument vorkommt. Die Dokumentenfrequenz sagt dagegen aus, in wie vielen Dokumenten einer

Menge an Dokumenten ein Wort vorkommt. Diese Methode bestimmt die Bedeutung eines Begriffes in einem Dokument im Verhältnis zu allen anderen Begriffen in diesem Dokument und die Häufigkeit des Auftretens dieses Begriffes in allen Dokumenten. Je häufiger ein Begriff in einem Dokument vorkommt und je weniger in allen anderen Dokumenten vorkommt, desto aussagekräftiger und daher relevanter ist dieser Begriff für dieses Dokument. Die Gesamtheit aller Dokumente wird dabei als Dokumentenkorpus bezeichnet. Bei der Menge an Twitternachrichten stellt sich jedoch die Frage, was man als Dokumentenkorpus wählt, da die Gesamtmenge aller Nachrichten sehr groß ist und stetig wächst. Es wäre daher denkbar lediglich einen zeitlichen Teilausschnitt zu betrachten, beispielsweise eine Woche oder einen Monat.

Die Relevanz eines Begriffes tj für ein Dokument dj wird durch folgende Formel berechnet. N stellt hierbei die Größe des Korpus, also die Menge der Dokumente dar.

$$w_{\mathrm{TF-IDF}}(t_j, d_i) = w_{\mathrm{local}}(t_j, d_i) \cdot w_{\mathrm{global}}(t_j) = \mathrm{tf}(t_j, d_i) \cdot \log\left(\frac{N}{\mathrm{df}(t_j)}\right)$$

Der TF-IDF Wert berechnet sich also aus dem Produkt der Term Frequenz und der logarithmierten invertierten Dokument Frequenz. Wenn wir die TF-IDF Methode auf unser vorheriges Beispiel anwenden, sähe die Berechnung wie folgt aus. Unser Dokumentenkorpus N bestünde in diesem Fall aus den Dokumente A, B und C. Wir wollen nun die Relevanz des Begriffes tj=Tag für das Dokument di=A berechnen. Die Term Frequenz von „Tag“ in Dokument A ist eins, da der Begriff ein einziges Mal vorkommt. Die Dokumentfrequenz des Begriffes beläuft sich auf drei, da er in drei Dokumenten vorkommt. Zusammen mit der Korpusgröße von drei, berechnet sich der Wert folgendermaßen.

W-tf-idf (Tag, Dokument A) = 1* log(3/3) = 0

Wie bereits erwähnt, verwenden viele Klassifikationsverfahren Vektoren für die Berechnung von Ähnlichkeiten. Daher müssen die Einträge der Term-Document Matrix in ein Vektorraum- Modell übertragen werden. Jedes Dokument wird dabei zu einem eigenen Vektor mit den verschiedenen Fragmenten als Dimensionen. Die Vektoren aus unserem Beispiel würden wie folgt lauten.

Vektor Dokument A = (11000000)

Vektor Dokument B = (01111100)

Vektor Dokument C = (11000000)

Die hierdurch erzeugten Vektoren dienen Klassifikationsverfahren wie Support Vektor Maschinen und K-Nearest Neighbor nun als Datenbasis. Bei der Sentimentanalyse wird die Distanz der Vektoren eines ungelabelten Dokumentes mit den als positiv, negativ und neutral eingestuften Trainingsdaten gemessen. Das Dokument wird danach in die Gruppe mit der höchsten Ähnlichkeit eingeordnet.

Neben den zwei Möglichkeiten semantik- oder lernbasiert, ist auch eine Mischform aus beiden Techniken denkbar. Hierzu wird ein lernbasiertes Verfahren verwendet, das jedoch bei der Term Document Matrix nicht alle in den Dokumenten vorkommenden Wörter mit einbezieht, sondern nur diejenigen, die Auswirkungen auf die Meinungsrichtung besitzen. Es wird hierbei eine Matrix aus dem Opinion-Lexikon erstellt, bei der jedes Wort eine Dimension darstellt. Danach wird für die manuell erstellte Trainingsmenge gemessen, welche Ausprägungen sie in den Dimensionen der Matrix besitzen. Einfacher ausgedrückt, welche Wörter aus dem Opinion-Lexikon kommen in der Trainingsmenge vor. Folgende Tabelle zeigt eine mögliche Term-Document-Matrix, die einen Teilausschnitt aller Dimensionen darstellt.

	gut	besser	am besten	schlecht	schlechter	am schlechtesten
Dokument A	1	1	0	0	0	0
Dokument B	0	1	1	1	1	1
Dokument C	0	1	0	0	0	0

Jedes zu klassifizierende Dokument wird nun nicht mehr nur anhand seiner meinungtragenden Wörter bewertet, sondern auch durch seine Nähe zu den Klassen der Trainingsmenge. Bei einer derartigen Vorgehensweise muss jedoch darauf geachtet werden, dass positive und negative Wörter ungefähr im gleichen Verhältnis vorkommen. Sollte dies nicht der Fall sein und beispielsweise weitaus mehr positive als negative Wörter in der Matrix vorkommen, führt dies dazu, dass der Algorithmus eine weitaus größere Menge an positiven Wörtern in den Tweets findet und sie dementsprechend zuordnet. Dies ist einer der entscheidenden Nachteile eines Opinion Lexikons und viele der bestehenden Lexika leiden unter diesem Problem. Hierdurch gibt es oft einen Überschuss an positiven Wörtern.

4.4 SentiStrengh

Alle die hier beschriebenen Sentimentanalysekonzepte haben den Nachteil, dass es sehr aufwendig ist aus ihnen ein funktionierendes System zu erstellen. Semantikbasierte Systeme benötigen ein entsprechend großes Sentimentlexikon und Methoden zur Erkennung der Grammatik wie Verneinungen und Fragestellungen. Lernbasierte Verfahren benötigen eine große Menge an Trainingsdaten, um gute Ergebnisse liefern zu können. Die Erstellung eines hinreichend großen und an den Analysezweck angepassten Trainingskorpus ist jedoch sehr aufwendig. Insbesondere wenn nicht nur eine binäre Einteilung in positiv und negativ vorgenommen werden soll, sondern eine hierarchische Skalierung mit mehreren Klassen. Deshalb haben wir uns dafür entschieden, ein bereits bestehendes und evaluiertes Framework zu verwenden. Vorgabe für dieses Framework war, dass es mit kurzen Texten umgehen können musste und möglichst keine Trainingsdaten benötigen sollte. Desweiteren sollte es als Java Package verfügbar sein, um es in die bestehende Architektur integrieren zu können. Schließlich sollte es nicht nur eine binäre Einteilung zwischen positiv und negativ vornehmen, sondern auch eine numerische Skalierung zwischen sehr positiv und sehr negativ berechnen

können. Nach einiger Recherche fiel die Wahl auf das SentiStrengh[17] Framework, da es als einziges alle Vorgaben erfüllen konnte.

SentiStrengh ist ein lexikonbasiertes Sentimentanalyseframework, das von der University of Wolverhampton entwickelt wurde. In der jetzigen Fassung kann es lediglich englische Texte bewerten, allerdings können die entsprechenden Lexika sehr leicht für andere Sprachen angepasst werden. SentiStrengh erstellt zwei Werte, einen Positivwert und einen Negativwert. Der Positivwert reicht von +1 bis +5, der Negativwert von -1 bis -5. Bei einer Addition der beiden Werte ergibt sich somit eine Spanne von -4 für sehr negativ bis +4 für sehr positiv. SentiStrengh ist für die wissenschaftliche Forschung kostenlos, eine kommerzielle Lizenz kostet 1000 Britische Pfund.

Das Framework ist nicht nur in der Lage den Sentimentgrad für einzelne Wörter zu bestimmen, sondern deutet auch aussagenverstärkende Wörter wie „sehr“ oder „weniger“. Desweiteren erkennt es eine Vielzahl an Emoticons und kann Verneinungen identifizieren. Sein Lexikon umfasst 890 positive und negative Sentiment-Wörter mit einer Stärkeskalierung von eins bis fünf. SentiStrengh ist desweiteren in der Lage, falsche Grammatik und Rechtschreibung zu erkennen. Ausrufezeichen werden als Verstärkungswörter erkannt. Negative Äußerungen, die mit einem Fragezeichen enden, werden dagegen ignoriert, da sie eine Fragestellung und keine Äußerung darstellen. Sollten meinungstragende Wörter mehrfach Selbstlaute enthalten wie z.B. „Niiiice“ wird dies als Verstärkung des Wortes interpretiert. Einer der größten Vorteile von SentiStrengh ist, dass es Domänenunabhängig ist. Vor allem lernbasierte Systeme haben das Problem, dass sie nur für eine spezielle Sorte an Textmaterial beziehungsweise für eine Quelle entwickelt wurden. SentiStrengh kann dagegen mit beliebigen Quellen umgehen und muss nicht erst auf ein neues Anwendungsgebiet angepasst werden.

[17] http://sentistrength.wlv.ac.uk/

4.5 Evaluationsmaße

Für die Bestimmung der Klassifikationsgüte wird sehr gerne die Konfussionsmatrix verwendet. Diese vergleicht die Vorhersage des Klassifikators mit einem Vergleichsdatensatz, bei dem die Zielvariable als korrekt klassifiziert angenommen wird. Dabei werden vier Fälle unterschieden. True Positiv, True Negativ, False Positiv und False Negativ. Als True Positiv und True Negativ werden Werte angesehen, die der Klassifikator korrekt als positiv beziehungsweise als negativ bewertet hat. Als False Positiv und False Negativ werden Werte angesehen, die vom Klassifikator falsch eingeordnet worden sind. Entweder wurde ein wahrer Wert als falsch angesehen oder ein falscher Wert als wahr gekennzeichnet.

		Vorhersage des Klassifikators	
		1	0
Tatsächliche Klasse	1	Wahre Positive (TP)	Falsche Negative (FN)
	0	Falsche Positive (FP)	Wahre Negative (TN)

Je höher die True Positiv und True Negativ-Werte im Verhältnis zu den False Positiv und False Negativ-Werten sind, desto besser arbeitet der Klassifikator.

Ein weiteres Qualitätsmaß ist der F-Score Wert. Dieser aggregiert die bereits im Information Retrievalabschnitt beschriebenen Maße Precision und Recall zu einer Variablen. Dies ist sehr hilfreich, da jede dieser Angaben für sich selbst nicht viel aussagt. Wenn ein IR System alle Dokumente zurück liefert, liegt der Recall bei 100%. Falls das System nur ein Dokument zurückliefert und dieses relevant ist, liegt die Precision bei 100%. Beide Zustände sind jedoch alles andere als wünschenswert. Der F-Score Wert verbindet nun die beiden Werte durch das harmonische Mittel aus Precision und Recall mit folgender Formel.

Ohne Gewichtung:

$$\frac{2}{\frac{1}{\mathbf{precision}} + \frac{1}{\mathbf{recall}}} = \frac{2 \cdot \mathbf{precision} \cdot \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}}$$

Mit Gewichtung:

$$F_\alpha = \frac{1}{\alpha \frac{1}{\mathbf{precision}} + (1 - \alpha) \frac{1}{\mathbf{recall}}}$$

4.6 Evaluation der bestehenden Techniken

Wie bereits erwähnt, gibt es keine einheitliche Vorgehensweise bei der Durchführung einer Sentiment Analyse. Hierdurch unterscheiden sich die Analysemethoden und die erzielten Ergebnisse deutlich voneinander. Um einen Überblick über die Ergebnisse des derzeitigen Forschungsstands zu bekommen, werden hier verschiedene Analysemethoden mithilfe einiger Qualitätsmaße miteinander verglichen.

Xiaowen Ding u.a konnten mit ihrem ganzheitlichen Ansatz das Feature Based Opinion Mining mit einer Distanzfunktion zu verbinden, folgende Werte erreichen: Precision 0,92, Recall 0,91 und F-Score 0,91. Wurde die Distanzfunktion nicht berücksichtigt, verschlechterten sich die Werte: Precision 0,86, Recall 0,89 und F-Score 0,87 vgl. (Ding, et al., 2008).

Vipul Pandey und Krishnakumar Iyer haben in ihrer Untersuchung verschiedene Klassifikationsmethoden untersucht und sind zu folgendem Ergebnis gelangt siehe (Pandey & Iyer, 2009). Nachrichten wurden von ihnen als Bag of Words betrachtet. Die Reihenfolge der Wörter wurde also außer Acht gelassen. Die Forscher berichten, dass ihre Klassifikationsgüte sich erheblich erhöhte, wenn sie unklar zuordbare Nachrichten aus den Trainingsdaten entfernten.

Classifier	Avg. F1	Avg. Acc.	Max. Acc.
Naive Bayes	0,563	75,63%	84,18%
SVM (Linear)	0,557	80,2%	85,44%
SVM (Poly)	0,499	79,22%	83,37%

Um die Genauigkeit zu verbessern, wurden Oversampling und Overboosting Techniken angewendet. Ohne diese Techniken hatte der Algorithmus einen niedrigen Recall jedoch auch eine niedrige false positive Rate. Boosting hat die Aufgabe die Genauigkeit zu erhöhen, indem es verschiedene schwache Klassifizierer zu einem starken vereint. Oversampling wird dazu verwendet, um Duplikate der Minderheitsklasse zu erstellen um den Größenunterschied der Klasseninstanzen auszugleichen und sie somit vergleichbar zu machen. Im Gegensatz dazu wird beim Undersampling die Mehrheitsklasse gekürzt. Durch Oversampling, Undersampling und Boosting konnte in der Untersuchung von Pandley und Iyer ein Recall von 85% erreicht werden.

Thomas Scholz hat in seiner Untersuchung die Genauigkeit von Klassifikationsverfahren gegenüber den verschiedenen meinungstragenden Wortformen untersucht. Diese Untersuchung kam zu folgendem Ergebnis vgl. (Scholz, 2011).

Wortart	Klassifikationsverfahren	Genauigkeit
Adjektive		
	Support Vektor Maschine	80,81%
	Naive Bayes	68,34%

	k-Nearst-Neighbour	53,6%
Verben		
	Support Vektor Maschine	82,07%
	Naive Bayes	72,29%
	k-Nearst-Neighbour	56,05%
Adverbien		
	Support Vektor Maschine	75,61%
	Naive Bayes	66,08%
	k-Nearst-Neighbour	53,79%

Man kann erkennen, dass die Support Vektor Maschinen bei Adjektiven, Verben und Adverbien am besten abgeschnitten haben. Diese Aussage deckt sich mit den Erkenntnissen der wissenschaftlichen Literatur wie (Pang, et al., 2002), die Support Vektor Maschinen ebenfalls als beste Klassifikationsmethode für Sentiment-Analysen empfehlen. Pang u.a. geben desweiteren Unigramme als beste Klassifikationsbasis an, vor allem wenn diese mit Support Vektor Maschinen kombiniert werden.

Nipun Mehra u.a. konnten in ihrer Untersuchung mit Hilfe eines Maximum-Entropie-Klassifikators eine Genauigkeit von 85% bei der Einteilung von Film Review in positive und negative erreichen vgl. (Mehra, et al., 2002). Sie geben an, dass ihr System so lange gut funktionierte, wie stark positive und negative Kritiken in den Reviews geäußert wurden.

Die Macher des SentiStrengh-Packages geben für ihre Software folgende Erkennungsgenauigkeit an. Hierbei muss berücksichtigt werden, dass es sich nicht nur um eine Einteilung in Positiv- und Negativ-Dokumente handelt, sondern um eine Einteilung in fünf verschiedene Stärkegrade bzw. Stärkeklassen.

	Genauigkeit	Genauigkeit +/- 1 Klasse
Positive Wörter	60,60%	96,90%
Negative Wörter	72,80%	95,10%

Man kann erkennen, dass die Genauigkeit für die exakte Klasseneinteilung mit 60,6% und 72,8% nicht übermäßig hoch ausfällt. Wenn man die Klassenzugehörigkeit allerdings etwas ungenauer fasst und diese auf +/- eine Klasse ausweitet, ist die Übereinstimmungsrate sehr hoch. Die Forscher geben an, dass maschinelle Lernverfahren im Moment etwas bessere Ergebnisse erzielen würden als ihr System. Allerdings sind diese immer domänenabhängig. Ihrer Meinung nach ist eine automatisierte Erkennung von Meinungen im sozialen Web möglich und dies sogar ohne manuell erstellte (Supervised) Trainingsdaten.

„Automatic classification of sentiment strength is possible for the social web –even unsupervised!" (Thelwall, 2011).

5 System zur Vorhersage von Börsenkursen

Nachdem wir durch die Auswertung von Forschungsarbeiten bewiesen haben, dass durchaus eine Verbindung zwischen Meinungen und Aktienkurs bestehen kann und dass eine automatisierte Erkennung von Meinungen mit den derzeit verfügbaren Mitteln möglich ist, gilt es diese Aussagen nun analytisch zu überprüfen. Hierzu wurde ein System entwickelt, dass von Twitter einen Strom an Tweets empfängt, die Meinungsrichtung des Tweettextes bestimmt und die Ergebnisse in eine Datenbank schreibt, um später Aussagen über den Meinungsverlauf treffen zu können. Im Folgenden wird nun der Aufbau des Systems beschrieben. Danach werden die verwendeten Vorverarbeitungsschritte erläutert. Schließlich werden die Ergebnisse aus der Analyse präsentiert.

5.1 Aufbau des Systems

Zur Sammlung des Datenmaterials wurde die Twitter Streaming API genutzt. Hierbei handelt es sich um eine von Twitter zur Verfügung gestellte Schnittstelle, die es ermöglicht, einen Strom von Tweets zu empfangen. Zur Nutzung der API ist die Verwendung von Twitters oAuth[18] System notwendig. Dieses sieht vor, dass man einen Benutzer erstellt, sowie mit diesem eine Registrierung der Software, die man erstellen möchte, bei Twitter vornimmt. Nach der Registrierung erhält man vier Hashwerte, den consumerKey, den consumerSecret, den accessToken und den accessTokenSecret. Diese Werte müssen entweder in einer Textdatei gespeichert werden, die im selben Ordner wie das Streaming-Programm liegt, oder man fügt diese direkt im Quellcode ein. Nur dann erhält man mit der Twitter Streaming API Zugang zu den Daten. Für die Anbindung von Software an die API existieren verschiedene Software Packages in verschiedenen Programmiersprachen. Wir haben uns für die Verwendung des Java Frameworks „twitter4j“ entschieden, da es sehr leicht zu implementieren ist und eine Java basierte Datenbankanbindung bereits vorhanden war. Die hierdurch erlangten Daten werden nun in einer Oracle Datenbank gespeichert. Die Datenbank wurde dabei mithilfe des Java Database Connectivity (JDBC[19]) Frameworks angebunden.

Dieses System wurde dazu verwendet, um Tweets der Firma Google Inc. zu sammeln. Um die Menge der Nachrichten nicht vollkommen ausufern zu lassen, wurden lediglich Tweets berücksichtigt, die das Hashtag „#Google“ enthalten.

Für die Erkennung der Sprache wurde das Language-Detection[20] Framework für Java genutzt. Dieses Framework unterstützt 53 Sprachen und verwendet für die Erkennung einen naiven Bayes Filter. Die Entwickler geben für Language-Detection eine Precision von 99% an. Dieses Framework liest den Tweettext aus der Datenbank, versucht die Sprache zu erkennen, und schreibt das Ergebnis zurück in die Datenbank. Somit ist nach diesem Schritt, die Sprache jedes Tweets bekannt, vorausgesetzt die Sprache war eine der vom Language-Detection unterstützten Sprachen.

[18] Oauth ist ein offenes und standardisiertes Protokoll, das es einer Webanwendung (Service Provider) erlaubt, Ressourcen mit anderen Webanwendungen (Konsument) auszutauschen, ohne den Benutzernamen oder das Password des Service Providers preisgeben zu müssen.

[19] JDBC ist eine einheitliche Schnittstelle zur Anbindung von relationalen Datenbanken verschiedener Hersteller an Java. Jeder Datenbankhersteller stellt hierfür einen speziellen Treiber zur Verfügung. In unserem Fall ist dies der Oracle Java Database Connectivity (OJDBC) Treiber in der Version sechs.

[20] http://code.google.com/p/language-detection/

Das SentiStrengh Framework wurde in derselben Weise wie das Language-Detection Framework angebunden. Dieses liest den Tweettext aus der Datenbank, jedoch dieses Mal nur jene Nachrichten, die das Language-Detection System als Englisch deklariert hat. Für jeden Tweettext liefert SentiStrengh nun einen Wert an Positivität und einen Wert an Negativität zurück. Diese beiden Werte werden danach zurück in die Datenbank geschrieben. Um die Sentimentwerte auf einer Zeitachse abzubilden, wurden gewöhnliche SQL-Anfragen an die Datenbank gestellt. Hierbei wurden die Sentimentwerte für jeden Tag summiert. Einen Überblick über die Struktur des Systems gibt Abbildung 11.

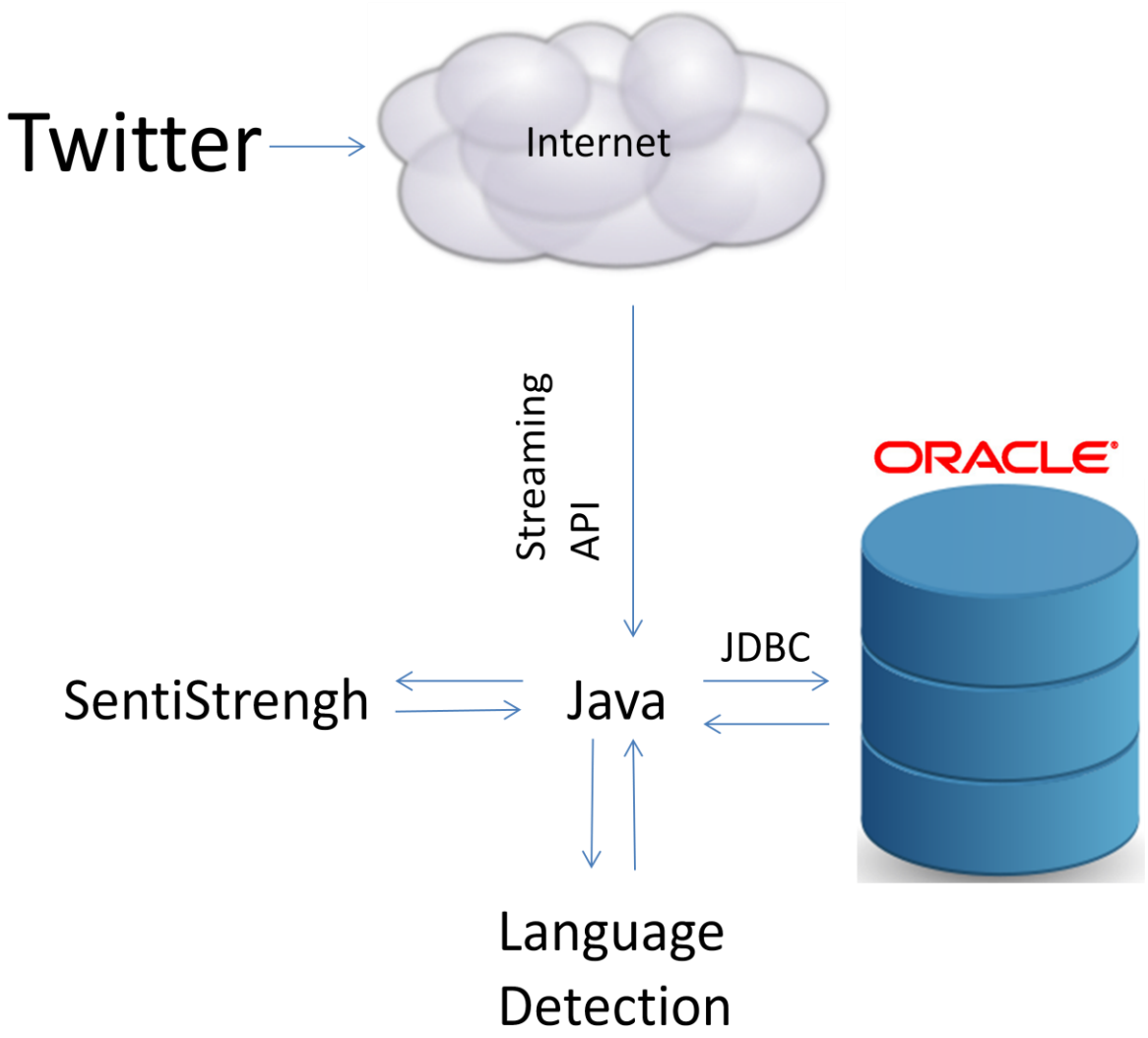

Abbildung 11: Aufbau des Systems

5.2 Vorverarbeitung

Der von Twitter stammende Rohtext der Nachrichten muss vor der eigentlichen Analyse aufbereitet werden, da er Elemente enthält, die entweder nutzlos sind, oder die Ergebnisse der Sprach- und Sentimenterkennung verfälschen würden. Die Vorverarbeitung erfolgt dabei in Java mit der Hilfe von regulären Ausdrücken[21]. Der ursprüngliche Text der Datenbank wird dabei nicht verändert. Stattdessen wird der Tweettext lediglich in Java gekürzt, bevor er dem

[21] Reguläre Ausdrücke sind Zeichen oder Zeichenketten, die es erlauben, durch syntaktische Regeln eine gesuchte Zeichenfolge zu beschreiben. Sie eigenen sich hierdurch sehr gut zur Suche bestimmter Textelemente.

Spracherkennungssystem und der Sentimenterkennung übergeben wird. Diese Form der Bereinigung wurde gewählt, da für die korrekte Spracherkennung die Hashtags aus dem Text entfernt werden müssen. Diese enthalten meist englische Wörter und würden somit die Erkennung der Sprache verfälschen. Für die Sentimenterkennung sind Hashtags allerdings sehr wichtig, da sie sehr of meinungstragende Worte enthalten.

Aus den in Abschnitt 4.1 vorgestellten Vorverarbeitungsschritten wurden in der Implementierung nun folgende ausgewählt.

1. Das Entfernen von Benutzernamen:

Benutzernamen bieten keinen Mehrwert für die Erkennung von Meinungen. Benutzernamen geben lediglich an, dass eine Nachricht an einen speziellen User gerichtet ist, dies hat jedoch keinen Einfluss auf die Meinungsrichtung.

2. Das Entfernen von Links:

Links bieten desweiteren für sich selbst keine Information, sie stellen nur eine Verbindung zu anderen Informationen her. Auf den Versuch diese Informationen zu berücksichtigen, wurde in dieser Arbeit verzichtet. Ein derartiges Vorgehen ist in der bisherigen Forschung noch nicht umgesetzt worden. Für die Einbindung der verlinkten Daten müssten die verlinkten Seiten analysiert werden. Dies ist jedoch nicht trivial, da der Link sich meist nicht auf alle auf der Webseite vorhandenen Informationen bezieht, sondern lediglich auf einen Teilausschnitt davon. Um eine maschinelle Verarbeitung dieses Teilausschnittes zu ermöglichen, müsste ein Algorithmus erkennen, auf welchen Abschnitt sich ein Tweet bezieht und lediglich dieses berücksichtigen. Da jedoch kein festes Identifikationsmerkmal existiert, das diese Teilausschnitte markiert, wäre es für einen Algorithmus sehr schwer, diesen zu erkennen. Für eine korrekte Zuordnung, müsste der Tweettext interpretiert werden und eine logische Zuordnung getroffen werden. Zu einer derartigen Interpretationsfähigkeit ist momentan allerdings nur der Mensch fähig.

3. Das Entfernen von nicht englischen Tweets:

Für die Sentimenterkennung werden alle Tweets die nicht in Englisch verfasst sind gelöscht, da das SentiStrengh Package lediglich mit englischen Texten umgehen kann.

Verzichtet wurde hingegen auf die Umformung auf folgende Punkte.

1. Das Entfernen von Abkürzungen:

Um Abkürzungen in ihre ungekürzte Form (z.B. wknd zu weekend) zu überführen, wäre eine Ableitungsregel für jedes einzelne Wort notwendig. Bei der Menge an möglichen Abkürzungen stände der Arbeitsaufwand in keinem Verhältnis mehr zum möglichen Nutzen. Abgesehen davon, besitzt das SentiStrengh Package bereits eine Funktion, die für viele Wörter eine falsche Schreibweise ausgleichen kann.

2. Das entfernen von SPAM:

SPAM ist maschinell nur schwer zu erkennen. Es existieren zwar eine Reihe von Wörtern, die einen kleinen Teil von SPAM-Nachrichten nahezu eindeutig identifizieren können, allerdings wird damit nur eine kleine Teilmenge aller möglichen SPAM-Nachrichten abgedeckt.

Was Retweets anbelangt, soll in dieser Untersuchung einmal eine Analyse mit und eine Analyse ohne Retweets erfolgen. Dies geschieht aus folgenden Gründen. Das Retweeten bedeutet, dass jemand die Nachricht für so bedeutsam angesehen hat, dass er sie an seinen Followerkreis weitergeleitet hat. Somit hat die Nachricht eine größere Anzahl an Personen erreicht und hat auf diese Weise auch mehr Einfluss bekommen. Es macht daher Sinn sie nicht aus der Analysemenge herauszunehmen. Bei der großen Anzahl an Retweets könnte das Entfernen von Retweets jedoch durchaus einen Unterschied in der Gesamtwertung auslösen. Insgesamt enthielt die Datenmenge 19,89% (68907) Retweets. Da solche Tweets durch die Anpassung der selektierenden SQL-Anfrage sehr leicht ausgeschlossen werden können, wird ein zweiter Test ohne Retweets erfolgen.

Abbildung 12 verdeutlich noch einmal die einzelnen Schritte der Vorverarbeitung.

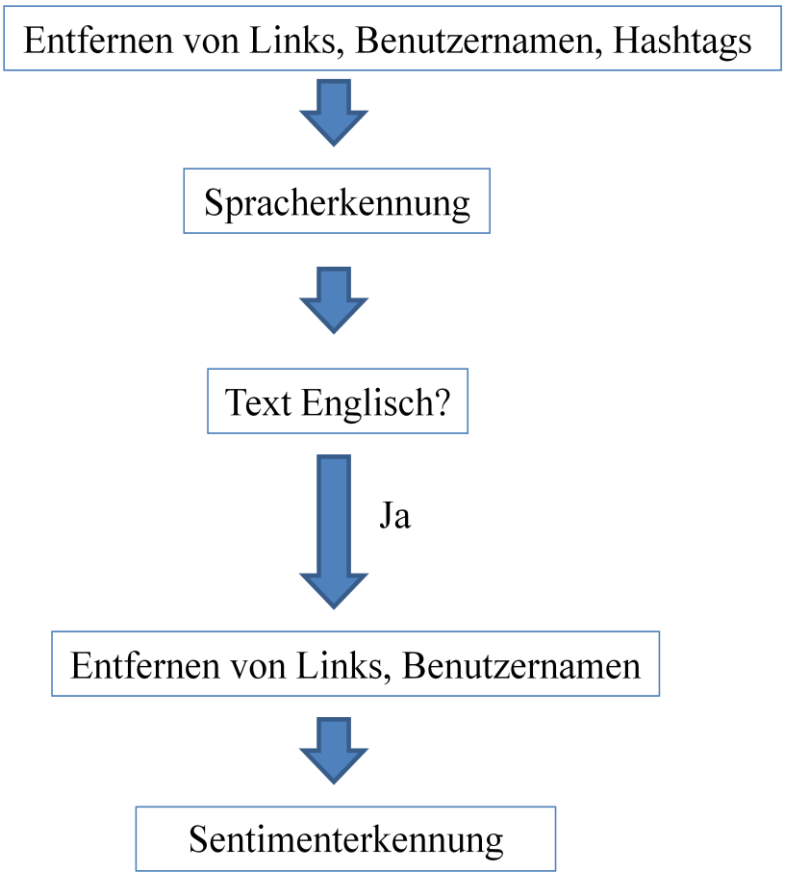

Abbildung 12: Vorverarbeitungsschritte

5.3 Analyse des Datenmaterials

Bevor wir mit der Auswertung der Sentimentdaten beginnen, soll ein Überblick über die Eigenschaften des verwendeten Datenmaterials erfolgen. Hierdurch sollen die Aussagen der wissenschaftlichen Literatur, was die Nachrichtenmenge, die Nutzereigenschaften und die verwendeten Sprachen anbelangt, überprüft werden.

5.3.1 Deskriptive Analyse

Die Trainingsdaten wurden im Zeitraum vom einem Monat vom 15.02.2012 bis zum 14.03.2012 erfasst. In dieser Zeitspanne wurden insgesamt 273294 Tweets gesammelt. Abbildung 13 zeigt die Schwankung der Nachrichtenmenge über den Analysezeitraum. Man kann erkennen, dass sich ein wellenförmiges Muster ergibt. Die Nachrichtenmenge steigt dabei in den meisten Fällen innerhalb der Woche immer weiter an, bis sie freitags oder samstags ihren Höhepunkt erreicht. Am Wochenende fällt sie dann stark ab und erreicht sonntags oder montags ihren Tiefpunkt bevor sie wieder anfängt zu steigen. Ferner lässt sich erkennen, dass die Nachrichtenmenge weiterhin ansteigt. Die Nutzung von Twitter scheint somit weiterhin zu steigen.

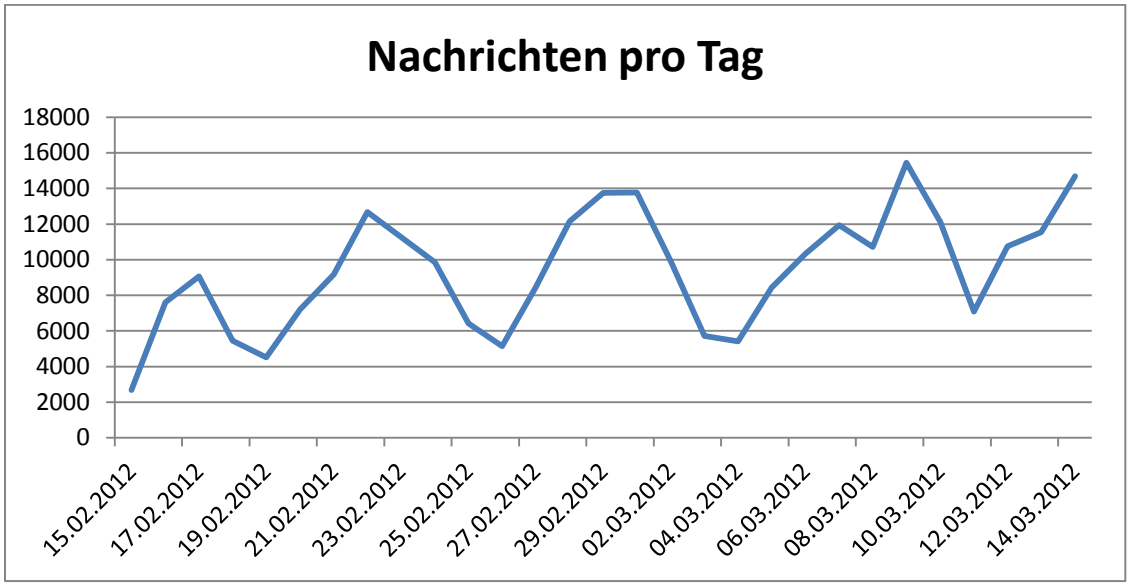

Abbildung 13: Nachrichtenmenge pro Tag im Betrachtungszeitraum

Kommen wir nun zur Analyse der Nutzer. Die Aussagen der wissenschaftlichen Literatur besagen hierzu, dass die Nachrichtenmenge sehr ungleich verteilt ist, also dass einige User sehr viele Nachrichten schreiben, während die meisten User nur sehr wenige Nachrichten verfassen.

Die Nachrichten unserer Untersuchung stammen von insgesamt 103989 verschiedenen Usern. Der User mit den meisten Nachrichten schrieb hierbei insgesamt 4988 Nachrichten, was bezogen auf den Betrachtungszeitraum von 29 Tagen eine Menge von durchschnittlich 172 Nachrichten pro Tag entspricht. Der Anteil der User, die nicht mehr als eine Nachricht verfasst haben, liegt dagegen bei 77% (80734).

Folgende Tabelle zeigt die Anzahl an Usern die eine gewisse Nachrichtenmenge verfasst haben. Das Intervall gibt an, in welchem Bereich die Nachrichtenmenge lag.

Intervall: Anzahl Nachrichten		
Von	Bis	Anzahl User
1000		24
500	1000	31
250	1000	45
100	250	86
50	100	115
25	50	352
10	25	1099
5	10	2123

Wie hoch ist jedoch die Menge, die einzelne User produzieren? Abbildung 14 zeigt die Nachrichtenmenge, die von den 61 produktivsten Usern erzeugt wurde. Es ist zu sehen, dass der Produktivste User Namens „akb_blog_googleplus“ mit 4988 Nachrichten einen sehr großen Vorsprung vor dem zweitplazierten User Namens „Hot Trend Now“ mit 2844 Nachrichten hat. Ab dem siebtplazierten User stellt sich dann ein linearer Abstieg in der Nachrichtenmenge ein.

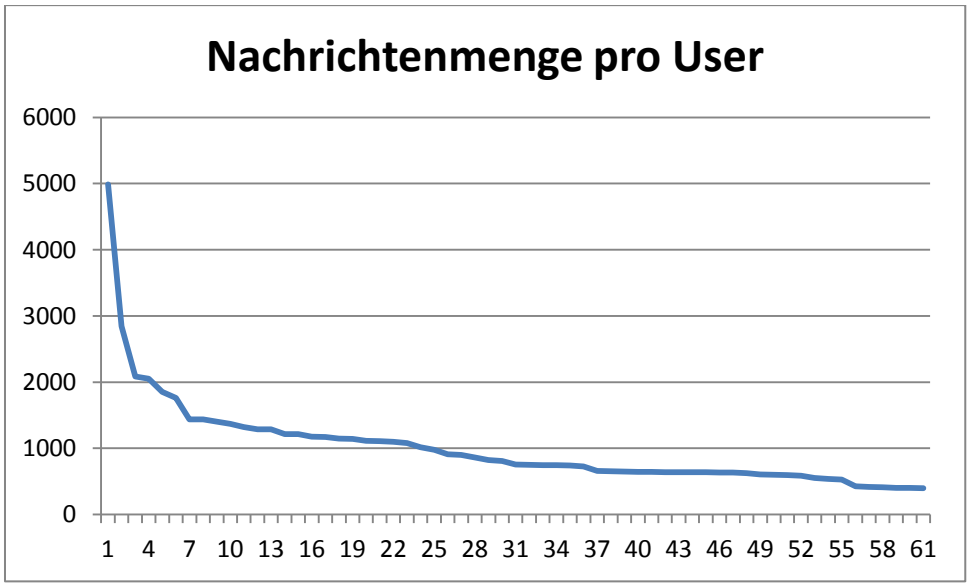

Abbildung 14: Nachrichtenmenge der 61 am häufigsten verkommenden User

In Abbildung 15 und 16 ist zu sehen, wie viele User eine gewisse Nachrichtenmenge verfasst haben. Die X-Achse zeigt die Nachrichtenmenge, die Y-Achse die Anzahl an Usern, die diese Menge an Nachrichten verfasst haben. Man kann erkennen, dass der größte Teil der User lediglich eine Nachricht verfasst hat. Danach fällt die Anzahl an Usern die mehr als eine Nachricht verfasst haben, sehr stark ab. Insgesamt lässt sich somit bestätigen, dass der Großteil der User sehr wenige Nachrichten verfasst, während eine kleine Anzahl an Usern eine sehr große Menge erzeugt.

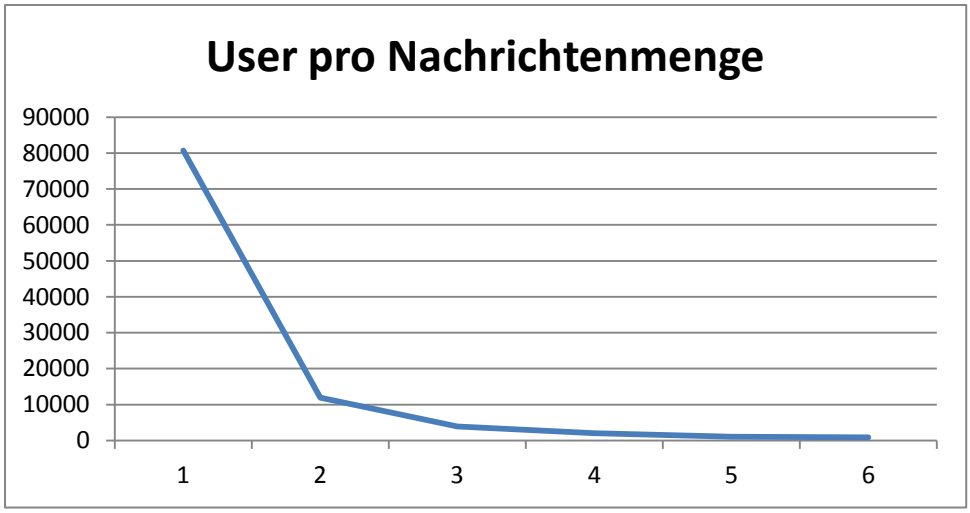

Abbildung 15: User pro Nachrichtenmenge – 1 bis 6 Nachrichten

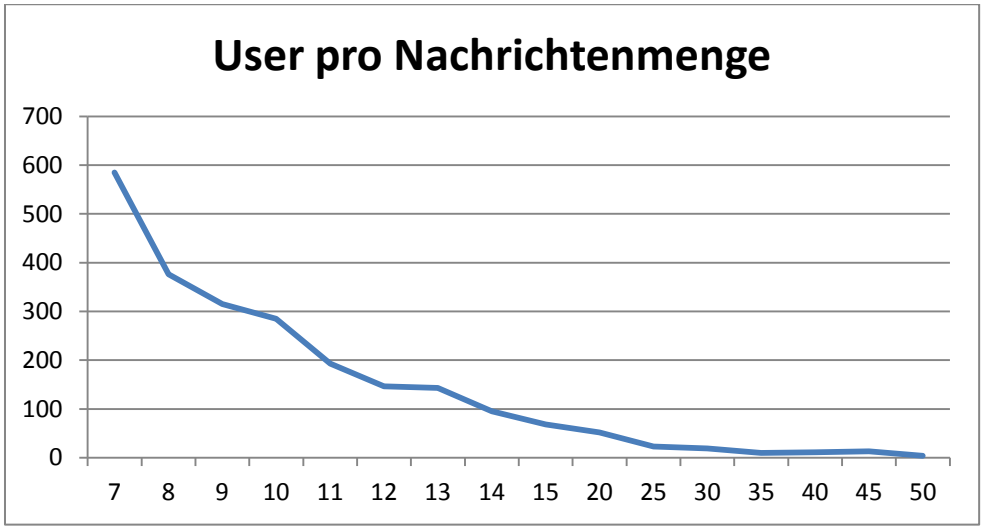

Abbildung 16: User pro Nachrichtenmenge – 7 bis 50 Nachrichten

Einige weitere Fakten zur Datenmenge. Insgesamt enthalten 74% (203268) aller Tweets Verlinkungen auf andere Webseiten. Lediglich 3,6% (9932) aller Nachrichten waren Antworten. Die höchste Anzahl an Followern bei einem User beläuft sich auf 3.288.175.

Die am häufigsten vorkommende Sprache ist wie zu erwarten Englisch mit 53,1% (145341). Um die Fehlerquote bei der Spracherkennung zu minimieren, wurden Links, Hashtags und Benutzernamen vor der Sprachanalyse entfernt. Vor allem die Entfernung von Hashtags war notwendig, da es viele Tweets gibt, die mehr Hashtags enthalten als Fließtext, wie folgender Beispieltweet zeigt.

#FB , #twitter ,, #BLOG , #Lockerz , #Google+ ,, #Frenlist .., smua ny aku punya ^_^/ hohoooo....

Derartige Nachrichten werden von Spracherkennungssystemen meist fälschlicherweise als englische Nachrichten interpretiert, da die meisten Hashtags englische Wörter enthalten. Neben der Mischung aus Hashtags und Fließtext gibt es auch Nachrichten, die ausschließlich aus Links und Hashtags bestehen wie folgender Tweet zeigt.

http://t.co/4jFyuuv0 #like please! #facebook #google #megusta #retweet #socialmedia #swagg #fail #meme #troll #mother #homero #simpson

Derartige Tweets werden lediglich dazu verwendet, um die in ihnen vorhanden Links zu verbreiten. Die große Menge an Hashtags wird dann dazu genutzt, um in möglichst vielen Listen aufzutauchen, die nach diesen Hashtags ihre Nachrichten filtern.

Wenn man nun die Anteile der zehn am häufigsten vorkommenden Sprachen darstellt, kommt man auf die in Abbildung 17 zu sehende Verteilung.

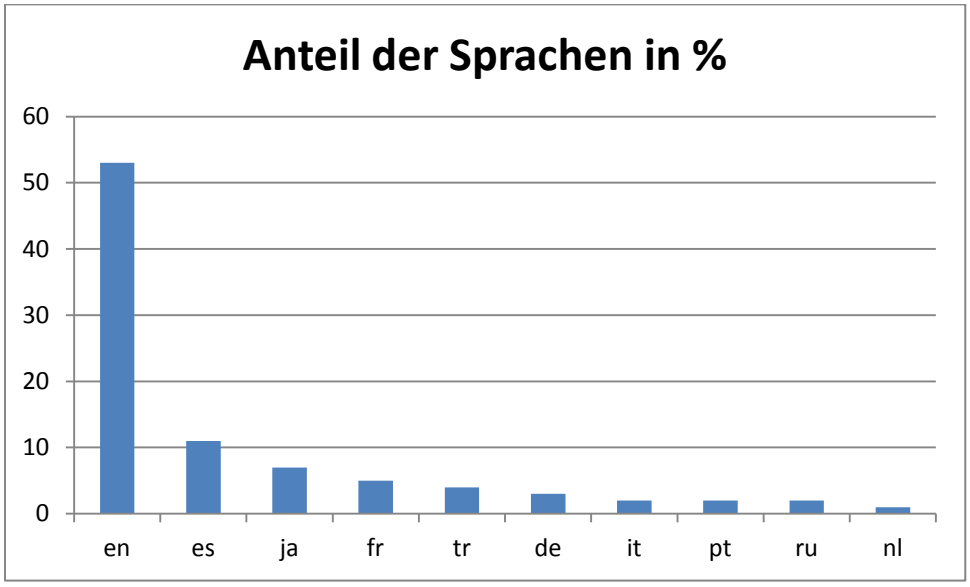

Abbildung 17: Anteil der Sprachen in den Daten

Wenn man diese Verteilung mit den Daten von (Hong, et al., 2011) vergleicht, die Englisch mit 51,1%, Japanisch mit 19,1% und Portugiesisch mit 9,6% angeben, stimmt lediglich die Anzahl englischer Tweets mit unserer Untersuchung in etwa überein. Der Anteil an japanischen Tweets lag in unserer Untersuchung bei 7%, der Anteil an portugiesischen Tweets bei lediglich 2%. Es verstärkt sich somit die Vermutung, dass es sich bei dem hohen Anteil von portugiesischen Tweets in der Untersuchung von Hong et al um einen Klassifikationsfehler handelt. Es ist dabei zu vermuten, dass es bei deren Analyse zu sehr vielen Verwechslungen zwischen spanischen und portugiesischen Nachrichten kam.

Um zu testen wie viele Tweets eine Meinung gegenüber Google beinhalten, wurden 100 Tweets manuell überprüft. 17 davon hatten eine positive Meinung und neun eine negative gegenüber Google. In dieser nicht repräsentativen Stichprobe hatten somit 26% aller englischen Tweets, die das Hashtag #Google enthalten, auch eine Meinung bezüglich Google. Hiermit währen in etwa ein Viertel der vorhandenen Nachrichten für eine Analyse geeignet. Folgende drei Beispiel-Tweets zeigen entsprechende Aussagen.

i always be the one who slow to know something, like this new #google #chrome features :(oh i love it

Remember #Google's old motto "Don't Be Evil"? Those were the days. Now it's "Fuck it & fuck you, we're huge & rich & we'll do what we want"

ok, it's official: #google #chrome is a fuckin #epicfail!! I need to change my browser now: any suggestions?

Um zu testen wie mehrdeutig meinungstragende Wörter sein können, wurden 100 Tweets manuell überprüft, die das Wort „fail“ enthalten. Davon waren 73 negativ und lediglich 4 positiv gegenüber Google. Somit lässt sich bei diesem Wort mit sehr hoher Wahrscheinlichkeit sagen, dass es einen Tweet als negativ klassifiziert. Allerdings enthalten lediglich 567 von den 157955 englischsprachigen Tweets dieses Wort.

Die Datenmenge enthält ebenfalls eine nicht unerhebliche Menge an SPAM-Tweets. Die folgende Nachricht kommt im Korpus bereits sechs Mal in gleicher oder leicht abgewandelter Form vor.

Selling More Shoes with Multi-Channel Funnels #google #dvg

Als SPAM können wohl auch die Vielzahl an Verlinkungen auf pornografische Webseiten angesehen werden. Alleine der Begriff „Blonde“ ergab im Datenmaterial über 100 Tweets mit Links zu derartigen Seiten. Von 100 manuell überprüften Tweets waren insgesamt sieben SPAM-Nachrichten, der Großteil hiervon Verlinkungen auf pornografische Webseiten.

5.3.2 Meinungsanalyse

Kommen wir nun zu den Ergebnissen der Meinungsanalyse des SentiStrengh-Packages. Bevor wir jedoch Aussagen über die aggregierten Analysedaten treffen, sollten wir einzelne Zuordnungen manuell überprüfen um die Sinnhaftigkeit der Ergebnisse zu validieren. Wie bereits erwähnt, ordnet SentiStrengh dem Text eine Sentimentskalierung von -4 bis +4 zu. Von 100 manuell überprüften Tweets deren Wert unter null lag, waren lediglich 30 tatsächlich negativ. Von 100 manuell überprüften Tweets, deren Wert über null lag waren 45 wirklich positiv. Ein Problem hierbei ist, dass viele der Tweets, die einen leicht positiven oder negativen Wert hatten, im Grunde neutral waren. Um zu überprüfen, ob eine stärkere Wertung zu besseren Ergebnissen führt, wurden wiederum 100 Tweets manuell überprüft, deren Wert über 1 beziehungsweise unter -1 liegt. Dies führte zu folgendem Ergebnis. Von 100 als positiv gelabelten Nachrichten waren 67 tatsächlich positiv. Von 100 als negativ gelabelten Nachrichten waren tatsächlich 63 negativ. Man erkennt, dass die Fehlerrate geringer wird, wenn schwache Wertungen nicht berücksichtigt werden.

Bei der manuellen Überprüfung fiel ebenfalls auf, dass es ohne Kontextwissen sehr oft auch als Mensch sehr schwer ist, zu entscheiden, ob eine Nachricht positiv oder negativ gemeint ist. Folgende zwei Tweets sind beispielsweise nur sehr schwer einzuordnen.

@sophayyy_: okay I don't know where looe is haha!” // Gotcha!!))o" lol!... #google ? Haha!

Awesome: 'Work from your heart, and really, really care.' http://t.co/ZvrhiUJu #apple #google #hertzfeld

Der erste ist eine Antwort und ohne die davor stattgefundene Unterhaltung zu kennen, ist eine richtige Interpretation des Textes nicht möglich. Der zweite Tweet enthält sehr viele positive Wörter, allerdings keinen direkten Hinweis, worauf sich diese beziehen. Diese Informationen sind lediglich auf der verlinkten Seite zu finden.

Der folgende Tweet könnte sarkastisch gemeint sein, ohne eine Überprüfung des Links ist hier eine eindeutige Aussage allerdings wiederum nicht möglich.

#Google #Hot #News Because what do I love more than boy bands, nothing! _#mycrazyobsession_ http://t.co/bnegl1wc

Das größte Problem sind jedoch Tweets, die sich nicht auf Google beziehen, obwohl sie dessen Hashtag enthalten. Nicht nur, dass sie nichts mit Google zu tun haben, sie können desweiteren ebenfalls starke sentimenttragende Wörter enthalten und somit die Analyse verfälschen. Folgender Tweet wurde beispielsweise mit -3 klassifiziert.

#Google #News Michael Madsen Busted for Child Cruelty http://t.co/0XWmFPG0 #InstantFollowBack GTNews

Folgender Nachrichtentweet wurde sogar mit -4 gelabelt.

#Google #Hot #News Will Goldman Sachs' CEO survive Greg Smith's 'devastating' rant? http://t.co/9NYqO4XJ #InstantFolowBack

Diese beiden Tweets zeigen, dass das Hashtag #Google sehr häufig zur Verbreitung von Nachrichten verwendet wird. Um dieses Problem zu umgehen, könnte man zwar alle Tweets löschen, die einen Link enthalten, allerdings würde man damit 74% aller Tweets entfernen. Desweiteren gehören Tweets, die Links enthalten, zu den glaubwürdigsten Nachrichten.

Manche Tweets beziehen sich zwar auf Google, die meinungstragenden Wörter dagegen auf etwas anderes. Folgender Tweet wurde beispielsweise als negativ deklariert obwohl er gegenüber Google sehr positiv ist.

@ryanm237 uses Twittee for "informational purposes." Lame...there's something called #Google for that ;)

Trotz all dieser Probleme liefert SentiStrengh durchaus brauchbare Ergebnisse. Um dies nachzuweisen, werden nun im Folgenden jeweils drei Tweets zu jeder Meinungskategorie gezeigt. Von sehr negativ bis sehr positiv. Die gezeigten Nachrichten stellen natürlich nur einen Ausschnitt aus der Gesamtmenge dar, allerdings geben sie durchaus einen Überblick in die Arbeitsweise des Algorithmus.

Rating: -4

Whoa, scary! Google Caught Tracking Safari Users: What You Need to Know

This item cannot be installed in your device's country" Oh, #Google, I really hate you for this message. D-:

#google I really hate you right now, #KillYourself

Rating: -3

Going through the whole #libjingle compiling maze... again. #Google hates developers

Microsoft Accidentally Marked As Malicious Site: If you're a Windows users, ... #google #search

Don't be evil! My ass... #google

Rating: -2

#Microsoft And #Google Jointly Failed To Beat #Apple In Market Cap #issues #business

Lack of API in #Google+ still hurting my use of the service. I find myself on FaceBook, Twitter, and 4sq more often due to tool usage.

Still no #iPad app for #Google+. Shame on you, Mountain View.

Rating: -1

do not use Google Plus-it is a giant spam scam operation for Google..#googleplus..#google.#spam..

Iran blocks #Google search, #Gmail, #Youtube and more

A look at some of the challenges #Google faces: Could Google Be Its Own Google-Killer? #searchmarketing

Rating: 0

20% of #Google+ users are students.

Small Web Company in UK sets standards for... #SEO #Business #Google #Marketing

#Android #Google EA's The Sims FreePlay available today on Android #DhilipSiva

Rating: 1

Dear #Google ,U always understand me :p

Always google updates the view of Gmail and the above task bar :D :)) #Google

#Android #Google HTC One X and One S to debut at Mobile World Congress #DhilipSiva

Rating: 2

I love being able to make phone calls from Gmail! Sound quality its rally good too! #google #googlevoice #gmail

The #Google doodle is soooo cute ! go check it out guys

I have 26 #google #newsbadges in subjects I didn't even know I liked reading about.

Rating: 3

Really enjoying #google wallet! Good work team

#Google Goggles! This is actually pretty awesome!!! Way better than QR codes. Unfortunately, NOT for BlackBerry!! NNOO!

She loved the one artificial rose out of a dozen tulips with my photos as the leaves.. Thanx to #google..

Rating: 4

Absolutely loving #Google's Chrome Web Store #lovetheseapps

LOVE #GOOGLE+ so fucking awesome!

I am a total #Android fanboy. I want #Google and Android to rule the world lol :-) :-) :-)

Auffällig ist, je extremer die Meinungsrichtung ist, desto sicherer sind die Ergebnisse des Algorithmus. Dies gilt sowohl in positiver als auch in negativer Richtung. Dies deckt sich mit der Vermutung, dass die Mehrdeutigkeit eines meinungstragenden Wortes mit dessen Stärke abnimmt. Entsprechend starke Worte sind jedoch sehr selten anzutreffen wie die folgende Verteilung der Meinungseinteilungen in Abbildung 18 zeigt.

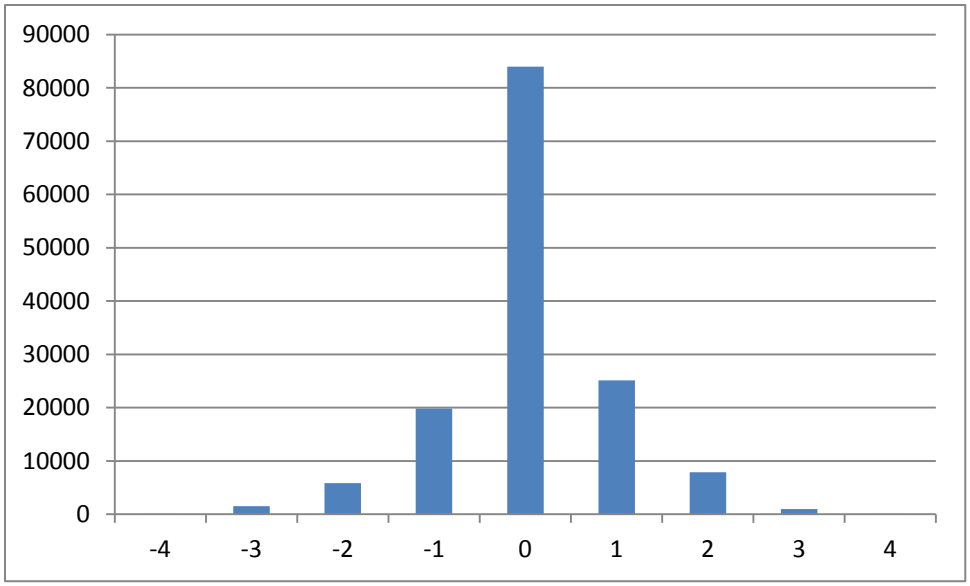

Abbildung 18: Bewertungsverteilung

Abbildung 18 zeigt die Verteilung der zugeordneten Meinungsrichtungen, mit der Bewertung auf der X-Achse und der Anzahl an korrespondierenden Tweets auf der Y-Achse. Auffällig ist, dass ein großer Teil der Tweets als neutral eingestuft wurde. Die Menge nicht neutraler Tweets beläuft sich auf 42% (61359), die von neutralen Tweets auf 57% (83982). Die Menge der geäußerten Meinungen nimmt dann sowohl im Positiven als auch im Negativen mit dem Grad an Stärke ab. Es existiert dabei ein kleiner Überschuss an positiven Nachrichten. Es wurden insgesamt 23,45% (34084) als positiv und 18% (27275) als negativ deklariert. Der größere Anteil an positiven Einteilungen könnte allerdings auch aus einer Ungleichverteilung von positiven und negativen Wörtern im verwendeten Opinion-Lexikon herrühren. Die Mengenverteilung stellt insgesamt eine Normalverteilung dar. Die Verteilung stimmt somit mit den Ergebnissen von (Godbole, et al., 2007) überein, die ebenfalls eine Normalverteilung messen konnten.

Bevor wir zum Vergleich der Sentimentdaten mit dem Börsenkurs kommen, soll ein Abgleich des Tweetaufkommens mit dem Aktienhandelsvolumen über den Betrachtungszeitraum erfolgen. Dies hat das Ziel, zu zeigen, ob ein logischer Zusammenhang zwischen beiden Werten besteht, beziehungsweise, ob beide ähnlichen gesellschaftlichen Faktoren unterworfen sind. Abbildung 19 zeigt die Menge an gehandelten Google-Aktien und die Anzahl geschriebener Tweets, die das Hashtag Google beinhalteten, im Betrachtungszeitraum. Beide Kurven weisen eine zyklische Schwankung mit einer großen Ähnlichkeit auf. Die Lücken im Aktienkurs rühren aus den Öffnungszeiten der Börse her, die lediglich zwischen Montag und Freitag geöffnet hat. Somit fehlen entsprechende Daten vom Wochenende.

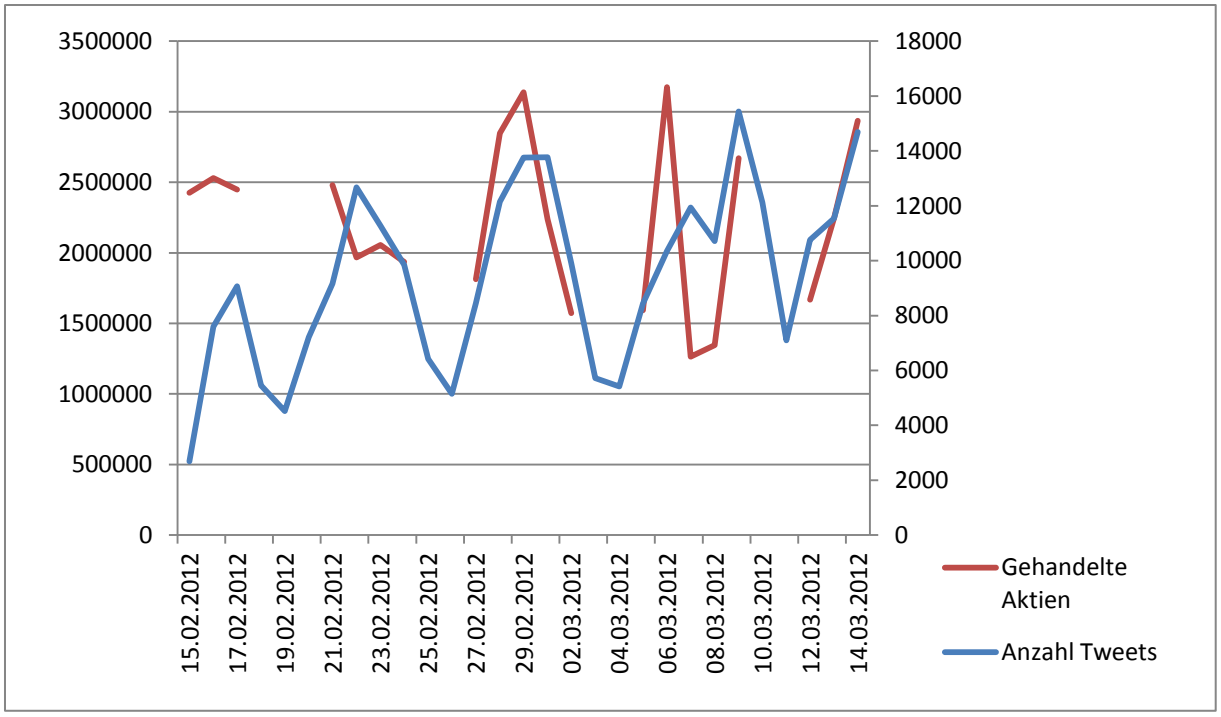

Abbildung 19: Anzahl Tweets + Handelsvolumen Google Aktien

Die große Ähnlichkeit der beiden Werte ist sehr überraschend, da zwischen beiden Werten eigentlich kein logischer Zusammenhang bestehen sollte. Eine derart große Ähnlichkeit weist jedoch darauf hin, dass beide ähnlichen gesellschaftlichen Strukturen unterworfen sind.

Kommen wir nun zum Abgleich der Sentimentdaten mit dem Aktienkurs und somit zur Hauptfragestellung dieser Arbeit. Um die Fragestellung zu beantworten, ob ein Zusammenhang besteht, werden die Sentimentkurve und die Aktienkurve übereinandergelegt und miteinander verglichen. Abbildung 20 zeigt in blau den täglichen Eröffnungskurs der Google Aktie an der New Yorker Börse im Zeitraum von Mitte Februar bis Ende März. Der Wert der Aktie kann anhand der Y-Achsen-Skalierung auf der linken Seite abgelesen werden. Die rote Kurve stellt die Sentimententwicklung über diesen Zeitraum dar. Der Wert der Sentimentkurve wurde für jeden Tag berechnet. Die vom SentiStrengh berechneten Werte wurden dabei durch die Anzahl an Nachrichten am jeweiligen Tag geteilt um eine bessere Vergleichbarkeit zu gewährleisten. Die Sentimentkurve wurde in Abbildung 20 um einen Tag nach vorne versetzt, da von Taytal & Komaragiri die Behauptung im Raum steht, dass der Sentimentkurs den Aktienkurs des nächsten Tages vorhersagen könnte. Wenn man beide Kurven betrachtet, kann man erkennen, dass im Zeitraum zwischen dem 15.02.2012 und 28.02.2012 eine gewisse Ähnlichkeit der Ausschläge besteht. Im Zeitraum vom 29.02.2012 bis 05.03.2012 weist die Sentimentkurve jedoch in die genaue Gegenrichtung der Aktienkurve.

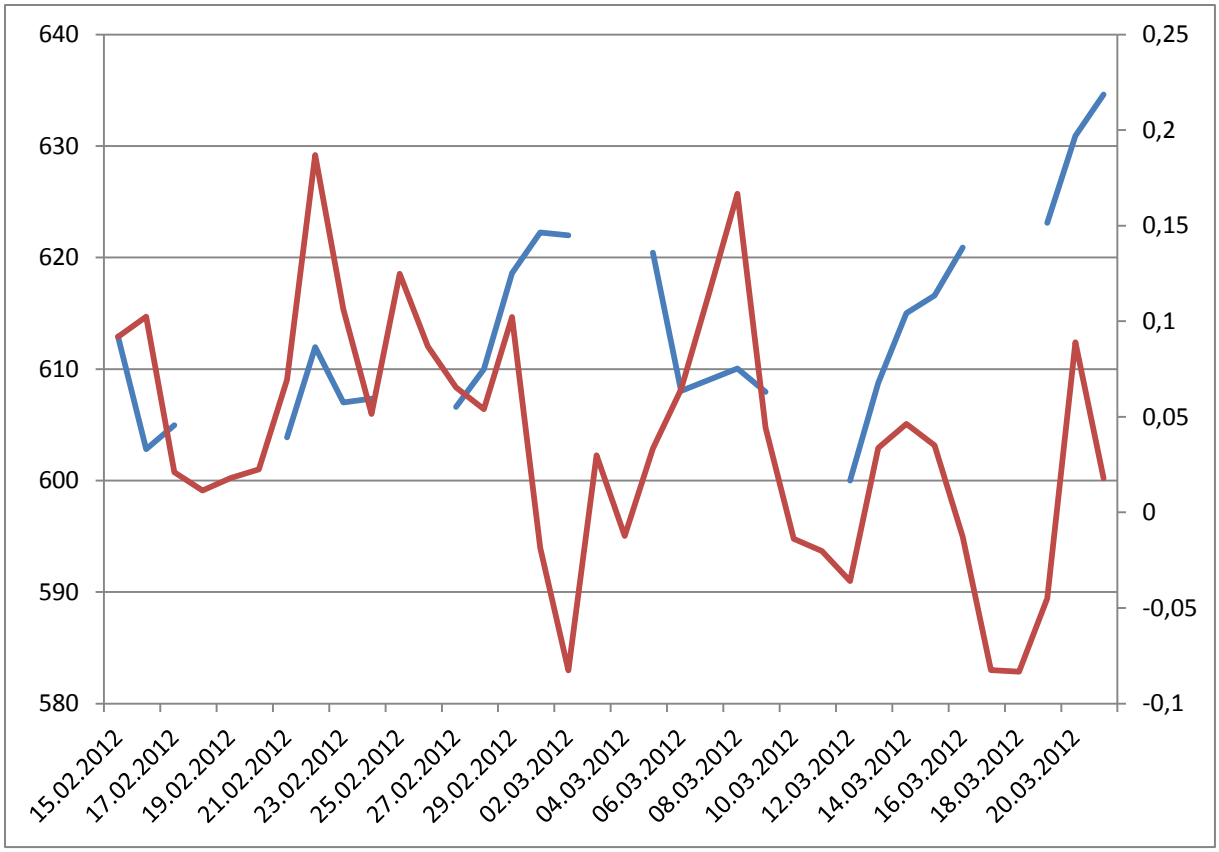

Abbildung 20: Aktien + Sentiment Kurs 1 Tag nach vorne verschoben

Von einer sehr hohen Übereinstimmung kann hier keine Rede sein, zu groß sind die Unterschiede in einzelnen Abschnitten. Eine gewisse Ähnlichkeit weisen die Kurven jedoch auf. Auffällig ist, dass der Sentimentkurs an den Wochenenden meist stark einbricht und danach wieder anfängt zu steigen.

Möglicherweise lässt sich die Übereinstimmung jedoch verbessern, wenn man von einer größeren zeitlichen Verschiebung des Sentimentkurses ausgeht. Anders ausgedrückt, ist es möglich, dass der Sentimentkurs nicht den Aktienkurs des nächsten Tages, sondern eines anderen in der Zukunft liegenden Zeitpunktes vorhersagen kann. Hierzu wurde der Sentimentkurs immer weiter nach vorne verschoben, bis eine möglichst große Übereinstimmung beobachtet werden konnte. Das beste Ergebnis konnte hierbei erzielt werden, wenn der Sentimentkurs acht Tage nach vorne verschoben wurde. Abbildung 21 zeigt das Ergebnis dieser Verschiebung. Eine derartige Verschiebung würde bedeuten, dass ein Kurs in der Sentimentkurve sich erst nach acht Tagen im Börsenkurs wiederspiegelt. Man kann erkennen, dass bis zum 16.03.12 eine deutliche Verbesserung in der Übereinstimmung erreicht wurde. Nach diesem Datum weisen die Kurven jedoch in die entgegengesetzte Richtung. Allerdings fangen beide Werte fast gleichzeitig ab dem 20.03.12 wieder an zu steigen.

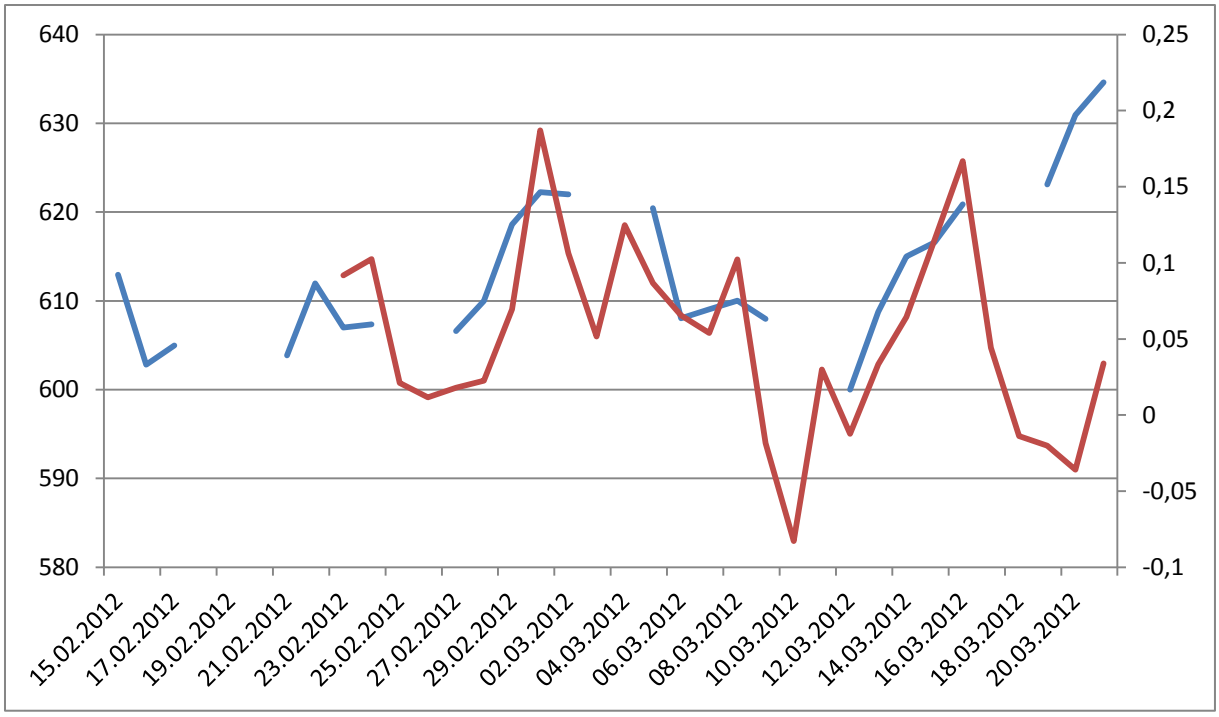

Abbildung 21: Aktienkurve + Sentimenkurve 8 Tage nach vorne verschoben

In den beiden vorangegangen Fällen wurden Retweets nicht entfernt und die gesamte Sentimentskalierung berücksichtigt. Wie bereits erwähnt, könnte das Entfernen von Retweets durch ihre große Anzahl durchaus einen Unterschied in der Sentimentaggregierung bewirken. Desweiteren hat die manuelle Sentimenteinstufung, wie bereits erwähnt, eine große Fehlerrate in den schwachskalierten Sentimenteinteilungen gezeigt. Aus diesen Gründen soll nun überprüft werden, ob das Entfernen von Retweets und schwachskalierten Tweets zu einer Verbesserung der Übereinstimmung zwischen Sentimentwerten und Aktienkurven führt. Zunächst soll der Einfluss von Retweets gezeigt werden. Abbildung 22 zeigt den Vergleich von Börsen- und Sentimentkurs ohne Retweets und ohne eine zeitliche Verschiebung.

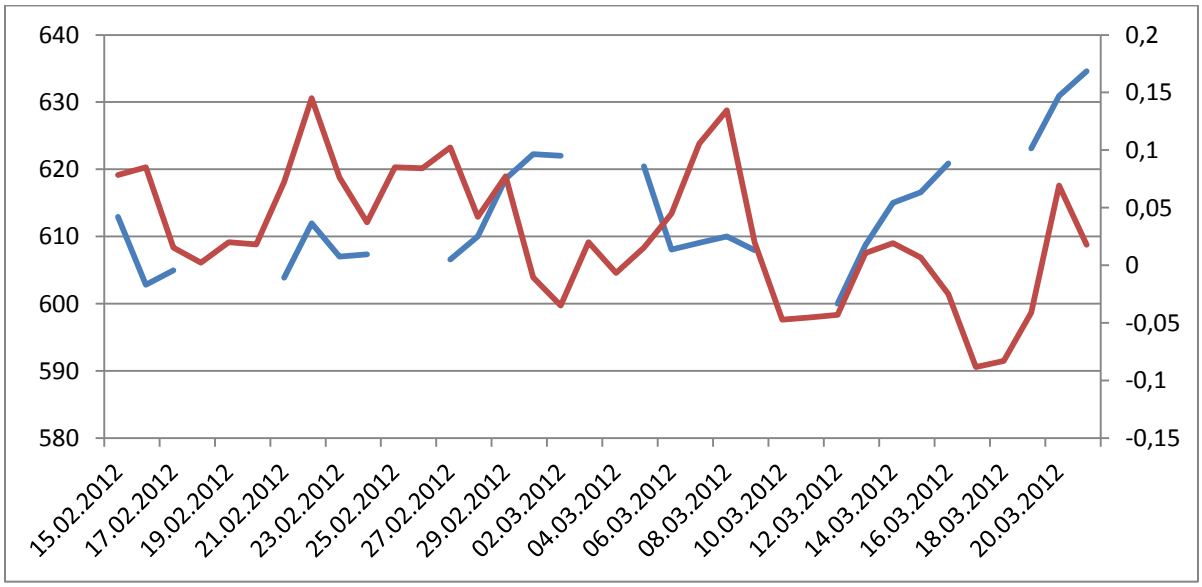

Abbildung 22: Sentimenkurs ohne Retweets

Es ist zu erkennen, dass der Sentimentkurs weniger extrem nach unten oder oben ausschlägt. Einen großen Unterschied im generellen Sentimentkurs konnte das Entfernen von Retweets jedoch nicht bewirken. Lediglich der Bereich vom 25.02.12 bis zum 28.02.12 unterscheidet sich leicht. Das Entfernen von Retweets konnte somit zu keiner signifikanten Verbesserung führen.

Kommen wir nun zur Auswirkung schwacher Sentimentbewertungen. Hierzu werden alle Tweets entfernt, deren Sentimentscore zwischen zwei und minus zwei liegt, also die Werte eins, null und minus eins. Abbildung 23 zeigt die Auswirkung dieser Maßnahme. Es ist ersichtlich, dass die Ausschläge der Sentimentkurve extremer ausfallen. Insgesamt verändern sich die entscheidenden Ausschläge jedoch nicht übermäßig.

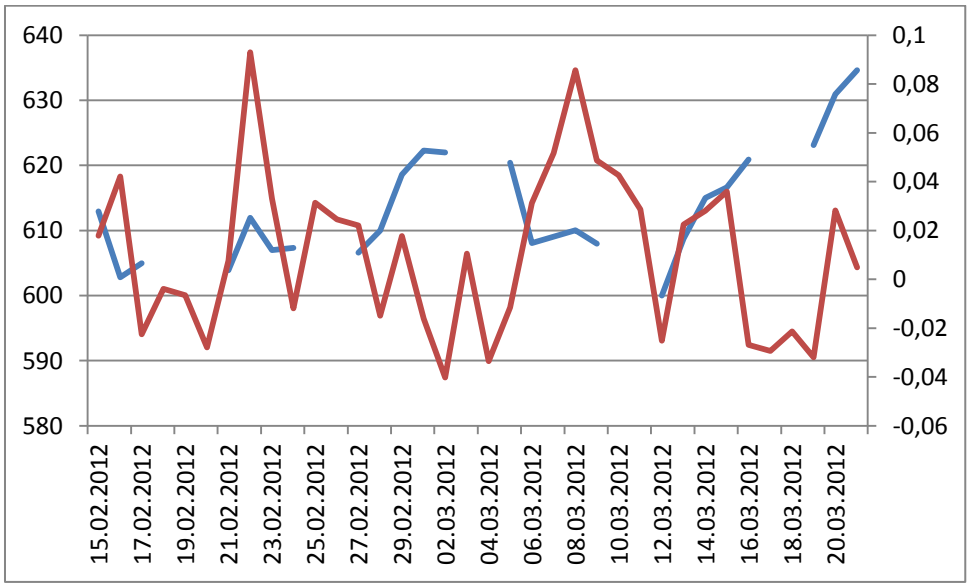

Abbildung 23: Sentimentkurs ohne schwache Bewertungen

Es stellt sich nun die Frage, ob sich trotz dieser Abweichungen ein Gewinn durch Investitionen anhand der Sentimentkurven erzielen lässt. Hierzu nehmen wir an, dass wir eine Aktie kaufen würden, wenn sich der Sentimentwert seit dem Vortag verbessert hat und die Aktie wieder verkaufen, wenn sich der Wert seit dem vorherigen Tag verschlechtert hat. Solange der Kurs steigt, wird die Aktie gehalten. Wenn wir diese Methode im Zeitraum vom 15.02.2012 bis zum 21.03.2012 anwenden, können wir einen Return on Investment (ROI) von 3,2795859 % erreichen. Hochgerechnet auf ein Jahr entspräche dies einem Zinssatz von 33,25%. Es sollte allerdings berücksichtigt werden, dass 58% des Gewinnes alleine durch die letzte Transaktion erzielt wurde.

Wenn diese Kaufstrategie lediglich auf den Aktienpreis angewendet wird, ergibt sich im Betrachtungszeitraum dagegen ein negativer ROI von -1,6314858%. Wenn man allerdings die Aktie am letzten Tag des Betrachtungszeitraums verkauft hätte, läge der ROI bei 4,01% und somit über dem der Sentimentanalyse. In beiden Fällen wird ein Großteil des Gewinns durch die starke Steigerung des Aktienkurses in den letzten neun Tagen erwirtschaftet. Folgende Tabelle zeigt alle Kaufentscheidungen.

Datum	Aktienkurs	Normalisierter Sentimentkurs	Kaufentscheidung nach Sentimentkurs	Kaufentscheidung nach Aktienkurs
15.02.2012	612,93	0,091801386		
16.02.2012	602,82	0,102381482	kaufen	
17.02.2012	604,97	0,020984507	verkaufen	kaufen
18.02.2012		0,011538462		
19.02.2012		0,018018018		
20.02.2012		0,022512151		
21.02.2012	603,87	0,06940483	kaufen	verkaufen
22.02.2012	611,96	0,187002096	verkaufen	kaufen
23.02.2012	607	0,106242312		verkaufen
24.02.2012	607,35	0,051494696		kaufen
25.02.2012		0,124784359		
26.02.2012		0,086743617		
27.02.2012	606,59	0,065393398		verkaufen
28.02.2012	610	0,053919348		kaufen
29.02.2012	618,6	0,102299881	kaufen	
01.03.2012	622,26	-0,018551689	verkaufen	
02.03.2012	622	-0,082722887		
03.03.2012		0,029945843		
04.03.2012		-0,012295082		
05.03.2012	620,43	0,033370412	kaufen	
06.03.2012	608,05	0,064422913		verkaufen
07.03.2012	609,05	0,114990689	verkaufen	kaufen
08.03.2012	610,04	0,166726683	kaufen	
09.03.2012	607,95	0,044257498	verkaufen	verkaufen
10.03.2012		-0,013895782		
11.03.2012		-0,020113583		
12.03.2012	600	-0,035921392		kaufen
13.03.2012	608,75	0,033814986	kaufen	
14.03.2012	615	0,046306355		
15.03.2012	616,6	0,03503225	verkaufen	
16.03.2012	620,89	-0,012833453		
17.03.2012		-0,082540339		
18.03.2012		-0,083370486		
19.03.2012	623,12	-0,044924844	kaufen	
20.03.2012	630,92	0,088916806		
21.03.2012	634,61	0,017777778	verkaufen	
			ROI: 0,03279586	ROI: -0,016314858

Wenn man die Sinnhaftigkeit der Kaufstrategien nun nicht nach dem Gesamtgewinn, sondern danach misst, wie viele Transaktionen einen Gewinn und wie viele einen Verlust erwirtschaftet haben, käme man auf folgende Ergebnisse: Die Kaufentscheidung aufgrund der Sentimentwerte konnte bei sieben Transaktionen fünf Mal einen Gewinn erzielen. Eine Kaufentscheidung aufgrund des Aktienkurses konnte dagegen bei keiner einzigen Transaktion einen Gewinn erzielen. Wenn man die Aktie am letzten Tag verkauft hätte, wäre es zu einer einzigen gewinnbringenden Transaktion gekommen. Ein wiederum sehr vielversprechendes Ergebnis für die Sentimentanalyse.

Wenn man den Aktienkurs mit dem Handelsvolumen der Google-Aktie vergleicht, kann man die Reaktion des Preises auf die gestiegene Nachfrage beobachten. Ein höheres Handelsvolumen resultiert fast immer auch in einem höheren Aktienkurs. Wie gewöhnlich führt hier eine höhere Nachfrage zu einem höheren Preis. Abbildung 24 zeigt beide Werte über den Zeitraum von vier Wochen.

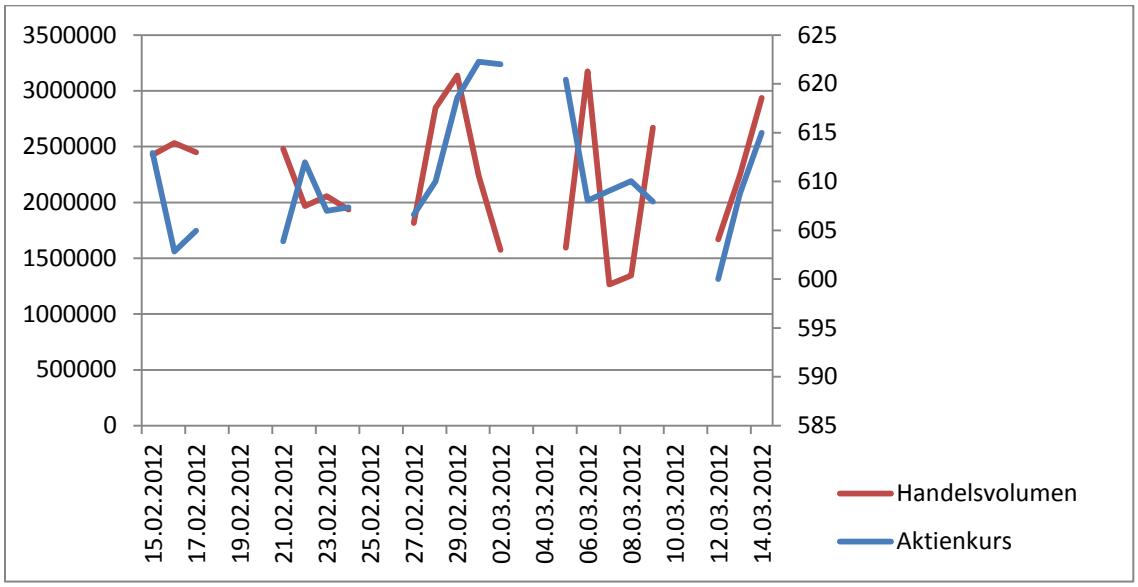

Abbildung 24: Vergleich Handelsvolumen - Aktienkurs

Da sich das Handelsvolumen der Google-Aktie und das Tweetvolumen aller Google Nachrichten sehr stark ähneln, stellt sich die Frage, ob das Tweetvolumen eine höhere Ähnlichkeit mit dem Aktienkurs besitzt, als der Sentimentkurs. Abbildung 25 zeigt die Überlagerung beider Kurven.

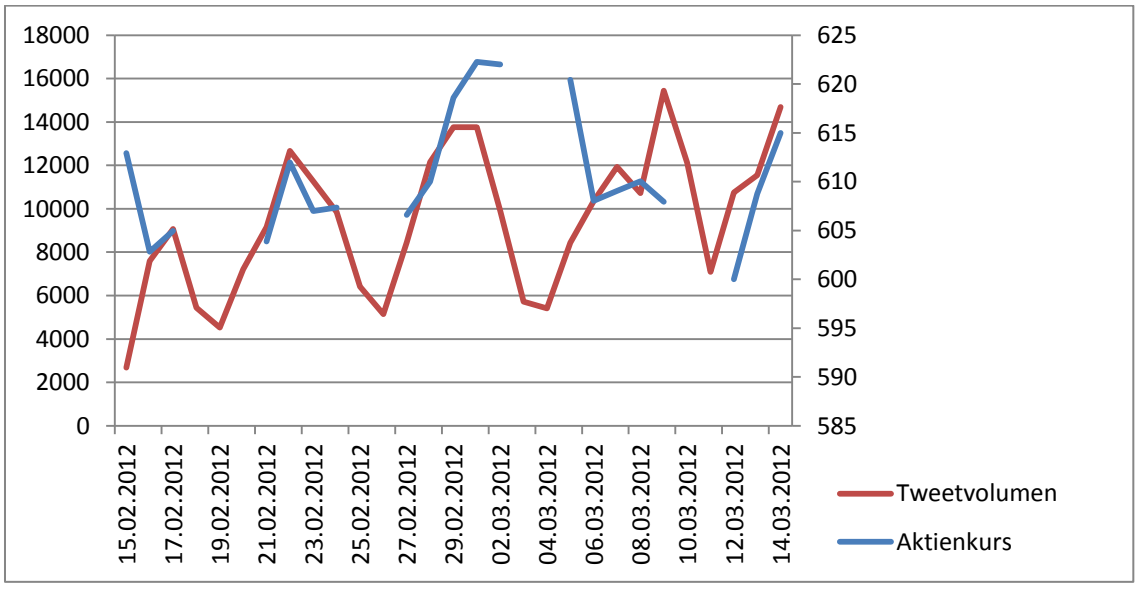

Abbildung 25: Vergleich Tweetvolumen - Aktienkurs

Wie beim Sentimentkurs gibt es auch beim Tweetvolumen einige Bereiche, die mit dem Aktienkurs übereinstimmen, einige sogar sehr genau. Allerdings gibt es eben auch Bereiche, in denen beide Werte stark voneinander abweichen.

6 Fazit & Ausblick

6.1 Daten

Die Isolierte Betrachtung einer Sentimentanalyse konnte den Aktienkurs im Betrachtungszeitraum dauerhaft nicht mit Sicherheit vorhersagen. In einigen Zeitabschnitten wiesen Sentiment- und Aktienkurs allerdings durchaus Übereinstimmungen auf. Interessant war auch die große Ähnlichkeit des Aktienhandelsvolumen und des Tweetaufkommens. Dies könnte darauf hindeuten, dass beide ähnlichen sozialen Strukturen unterworfen sind.

Für eine genauere Aussage wäre es allerdings notwendig gewesen, eine größere Zeitspanne zu betrachten. Eine wirklich verlässliche Aussage ließe sich sicherlich erst nach einigen Jahren der Analyse treffen. Dass eine einzige Quelle einen so vielschichtigen Index wie den Aktienkurs mit nahezu 100% prozentiger Sicherheit vorhersagen kann, war von vorneherein mehr als unglaubwürdig. Eine Frage, die durch den zeitlich begrenzten Rahmen dieser Arbeit jedoch noch unbeantwortet bleibt, ist, ob auf längere Sicht die Übereinstimmungen des Aktien- und Sentimentkurses die Abweichungen dieser beiden Werte übersteigen und somit zumindest eine gewisse Tendenz zu einer Ähnlichkeit bestände.

Die Verwendung einer Sentimentanalyse zur Investitionsentscheidung konnte dagegen im betrachteten Zeitraum sehr gute Ergebnisse liefern. Wiederum muss sich hierbei noch beweisen, dass diese guten Ergebnisse auch über einen längeren Zeitraum gehalten werden können. Ein ROI von 3,28% innerhalb von neun Wochen ist allerdings mehr als beachtlich, da er damit die regulären Gewinnmargen, die am Aktienmarkt normalerweise erwirtschaftet werden, um ein vielfaches übersteigt.

Die Auswertung der wissenschaftlichen Literatur konnte zeigen, dass Twitter eine sehr gute Datenquelle für die Untersuchung gesellschaftlicher Phänomene ist insbesondere für Marktforschung und Trendanalysen. Durch die breite gesellschaftliche Verbreitung von Twitter sind Untersuchungen dabei für fast alle gesellschaftlichen Phänomene möglich. Wie die Altersstruktur gezeigt hat, sind lediglich sehr junge und sehr alte Menschen unterdurchschnittlich vertreten. Analysen die, diese beiden Gruppen betreffen, sind somit wenig sinnvoll.

Wie gut dies im Einzelnen funktioniert, hängt allerdings stark von der verfügbaren Datenmenge ab. Ohne eine hinreichende Anzahl an Tweets zu einem Thema, einer Firma oder einem Produkt, ist eine Aussage aufgrund dieser Daten nicht repräsentativ und somit unbrauchbar. Desweiteren ist es derzeit nicht möglich in einer Menge von Tweets differenzierte Nutzergruppen herauszufiltern. Ein großes Problem ist hierbei, dass das Alter eines Users nicht verfügbar ist. Die Angaben zur geografischen Position sind desweiteren meist nur über den angegebenen Wohnort des Users bestimmbar. Um die Position des Wohnortes zu bestimmen, muss dieser mit einem Lexikon aller Städte und Dörfer abgeglichen werden. Durch diesen Mangel an speziellen Userdaten sind Aussagen zu einzelnen Altersgruppen nicht möglich. Aussagen aufgrund des geografischen Umfeldes sind dagegen zwar möglich, jedoch sehr aufwendig. Ein derartiger Sachverhalt limitiert den Einsatz von Twitter für Anwendungen, die eine speziell abgestimmte Stichprobe einer bestimmten Bevölkerungsgruppe benötigen. Trotz dieser Einschränkungen wird durch die bereits

bestehenden Erfolge bei der Analyse von Twitternachrichten dieser Forschungszweig in Zukunft wohl noch weiter Zulauf erhalten. Wie sehr hängt allerdings auch von der weiteren Entwicklung von Twitter ab. Derzeit steigen die Nutzer- und Nachrichtenzahlen weiter an, wie lange diese Entwicklung weitergeht ist momentan jedoch nicht abzusehen. Die wenigen Informationen, die die Analyse von Twitternachrichten erschweren, könnten von Twitter ohne großen Aufwand nachgeliefert werden. Angaben zum Alter könnten von Twitter beispielsweise problemlos abgefragt werden. Was die Lokalisierung anbelangt, wäre ein Annäherungswert über die IP Adresse des Users denkbar, wenn keine GPS Daten vorhanden sind. Dies würde zwar keine exakte Lokalisierung erlauben, allerdings würde es den Radius sehr stark eingrenzen. Ob Twitter derartige Änderungen plant, ist derzeit unbekannt. Allerdings könnten Firmen Druck auf Twitter ausüben, ihren Dienst um diese Informationen zu erweitern.

6.2 Methoden

Die Klassifikationsgüte des SentiStrengh-Packages bewies wie gut heutige Algorithmen die Meinungsrichtung bereits bestimmen können. Die Zukunft wird hier sicherlich noch exaktere Frameworks hervorbringen. Der nächste Evolutionsschritt stellt hier sicherlich die Anwendbarkeit auf viele Sprachen dar, um möglichst viele Tweets in die Analyse mit einbeziehen zu können. Eine gewisse Fehlerrate muss bei einem automatisierten Sprachverarbeitungssystem jedoch akzeptiert werden. Die Feinheiten der menschlichen Sprache sind zu vielschichtig, komplex und von ständigen Änderungen bestimmt, um ein perfektes Ergebnis erreichen zu können. Die Meinungsrichtung einer einzelnen Aussage kann somit nie absolut eindeutig bestimmt werden. Eine aggregierte Bewertung auf eine Menge von Aussagen kann allerdings durchaus sichere Ergebnisse liefern. Die verwendeten Lexika müssen allerdings im Zeitverlauf immer wieder angepasst werden, da sich die Bedeutung von Wörtern ändern kann vgl. (Burkert, 1996).

Mit dem steigenden Gebrauch von Twitter als Analysewerkzeug wird es wohl auch zu einem steigenden Missbrauch der Plattform zur Kontrolle, Manipulation und persönlichen Vorteilname kommen. Durch die Hilfe von Sentimentanalysen könnte man versuchen, die Einstellung einer Person zu gewissen Themen wie Politik oder Religion zu messen. Autoritäre Regime hätten mit einem derartigen Analysewerkzeug beispielsweise ein Mittel in der Hand, um unliebsame Regimekritiker aufzuspüren und zu verhaften. Aber auch in demokratischen Staaten könnten Firmen auf die Idee kommen, die Einstellung ihrer Mitarbeiter zu überwachen. Zum Beispiel, um ihre Einstellung zur eigenen Firma oder Vorgesetzten zu messen, vorausgesetzt diese würden sich darüber äußern. Man muss sich darüber im Klaren sein, dass Twitter eine öffentliche Plattform ist, und dass jeder die eigenen Nachrichten empfangen kann. Die meisten Menschen sind sich dessen bewusst, viele nutzen den Dienst allerdings zu unbedarft.

Sollte Twitter zukünftig vermehrt als Investitionsgrundlage für Aktienkäufe dienen, bestünde auch die Gefahr, dass einige versuchen durch das Verbreiten von Falschinformationen die Börsenmakler und somit auch den Aktienkurs zu beeinflussen. Im regulären Börsenbetrieb ist ein derartiges Vorgehen nicht unüblich. Um die Aussagen einer Twitteranalyse nicht zu verfälschen, müssten derartige Fehlinformationen erkannt werden. Ein Indikator hierfür könnte die Menge an Nachrichten sein die ein User verbreitet, da es viele Nachrichten bedarf,

um eine Aggregation über tausende von Nachrichten zu beeinflussen. Ein User müsste schon sehr viele falsche Nachrichten verbreiten um eine Wirkung zu erzielen. Vor allem ein schneller Wechsel eines Users von wenigen Nachrichten pro Tag zu sehr vielen, könnte auf einen Manipulationsversuch hindeuten. Gegen einem verteilten Angriff vieler Nutzer, ähnlich einer Distributed Denial of Service (DDOS[22]) Attacke, könnte man sich allerdings kaum wehren. Allerdings müsste ein Angreifer hierfür die Twitterkonten sehr vieler Nutzer unter seine Kontrolle bringen, was zwar nicht unmöglich wäre, allerdings dürfte es den Manipulationsversuch durchaus erschweren.

Auch wenn keine direkte Übereinstimmung zwischen Sentiment- und Aktienkurs bewiesen werden konnte, zeigte die Ausgereiftheit des SentiStrengh-Frameworks, dass automatisierte Sentimentanalysen bereits jetzt funktionieren. Bei der Macht und dem Einfluss den die Online-Community bereits heute hat, kann es sich ein Unternehmen nicht mehr leisten, die in sozialen Netzwerken wie Twitter geäußerten Meinungen zu ignorieren. Insbesondere da sich schlechte Publicity im Netz explosionsartig ausbreiten und somit den Ruf einer Firma oder eines Produktes sehr belasten kann. Es wird daher immer mehr ein erfolgsentscheidender Faktor, die Meinung der Netzuser zu kennen. Desweiteren ist es nicht nur möglich die Meinungen über die eigene Firma zu untersuchen, sondern auch die der Konkurrenz. Somit sind Vergleiche zwischen den Firmen und ihren Produkten und Dienstleistungen möglich. Folglich sind auch Rückschlüsse über das Image möglich, das zu einem nicht unwesentlichen Teil zum Unternehmenserfolg beiträgt. So lässt sich ebenfalls ein Markenmonitoring betreiben und Trends in der Entwicklung der eigenen und fremden Marken aufspüren. Auch neue Trends lassen sich somit erkennen. Sollte es einer Firma darüber hinaus gelingen, eine Person, die in einem sozialen Netzwerk wie Twitter sehr viele Zuhörer besitzt, für sich zu gewinnen, kann deren Meinungsmacht für sich genutzt werden. Aus all diesen Gründen ist es daher nur noch eine Frage der Zeit bis Sentimentanalysen großflächig Anwendung finden.

Im Hinblick auf die verwendeten Klassifikationsalgorihmen konnte sich bisher noch keiner als vollkommen überlegen zeigen. Die Methoden des maschinellen Lernens, die eine Zuordnung lediglich aufgrund von Ähnlichkeitsmodellen liefern, können zwar in einem auf sie abgestimmten Fall gute Ergebnisse liefern, allerdings sind sie nicht generalisierungsfähig. Der Hauptgrund dafür liegt darin, dass die Trainingsdaten eine starke Ähnlichkeit zu den Analysedaten haben müssen. Dieser Sachverhalt war auch der Grund dafür, dass wir uns in dieser Analyse zu einem semantikbasierten Verfahren entschieden haben, da eine genügend große Anzahl an bewerteten Twitter-Nachrichten nicht vorhanden war. Eine manuelle Erstellung einer hinreichend großen Trainingsmenge wäre sehr zeitaufwendig gewesen. Insbesondere da nicht nur eine binäre Einteilung in positiv und negativ von Nöten war, sondern eine numerische Skalierung mit vielen Klassen. Somit hätte für jeden Skalenwert eine eigene Trainingsmenge erstellt werden müssen. Lediglich eine Skalierung basierend auf dem Ähnlichkeitsmaß hätte sicherlich zu sehr ungenauen Ergebnissen geführt, da es die Ähnlichkeit zu einer Menge an Dokumenten getestet hätte, die keine einheitliche Meinungsstärke besitzt. Sollte eine Unterstützung von mehreren Sprachen von Nöten sein,

[22] Bei einer Denial of Service Attacke handelt es sich um die mutwillige Überlastung eines Rechners, meist eines Webservers, mit dem Ziel seine Dienste außer Funktion zu setzen. Der Begriff Distributed sagt dabei aus, dass es sich nicht um einen einzelnen Angreifer handelt, sondern um eine verteilte Attacke vieler Angreifer.

müsste desweiteren für jede Sprache eine eigene Trainingsmenge erstellt werden, was den Aufwand vervielfachen würde.

Ein semantik- bzw. lexikonbasierter Ansatz, wie der des genutzten SentiStrengh-Packages, kann dagegen weitaus vielseitiger eingesetzt werden. Um hierbei jedoch gute Ergebnisse zu erzielen, muss der Algorithmus neben dem reinen Vorkommen eines Wortes auch die Satzstellung und grammatikalischen Besonderheiten berücksichtigen. Anders lassen sich die schlechten Ergebnisse von (Ohana & Tierneyy, 2009) nicht erklären, die mit Hilfe des SentiWordNet lediglich eine Klassifikationsgenauigkeit von 65,85% erreichen konnten. Ein Sentiment Lexikon bietet desweiteren den Vorteil, dass es sich sehr leicht an neue Sprachen anpassen lässt. Es ist somit zu erwarten, dass lexikonbasierte Systeme wie SentiStrengh in absehbarer Zeit für eine Vielzahl an Sprachen angeboten werden.

Derzeit werden Sentimentanalysesysteme meist von Forschern und kleinen Softwarefirmen entwickelt. Wissenschaftler teilen ihre Erkenntnisse zwar in Form von Papern, allerdings halten sie den Quellcode ihrer Softwaresysteme meist geheim. Softwarefirmen haben desweiteren finanzielle Interessen an der Vermarktung ihrer Software und sind daher erst recht nicht gewillt ihre Erkenntnisse oder gar den Quellcode offen zu legen. Dieser Sachverhalt erschwert die Weiterentwicklung von Sentimenterkennungssystemen erheblich und schafft viele Insellösungen, die nicht voneinander profitieren können. Ferner fehlen einzelnen Forschergruppen und kleinen Softwarefirmen die nötigen Ressourcen um komplexe Systeme zu entwickeln. Wie sich die weitere Entwicklung vollzieht hängt wohl davon ab, wie sich der Markt bzw. die Nachfrage für derartige Software und somit der erzielbare Gewinn entwickelt. Sollten immer mehr Firmen das Potential maschineller Meinungsforschung erkennen, ist es sehr wahrscheinlich, dass die Nachfrage nach entsprechender Software sehr stark zunehmen wird. Spätestens dann werden sich wohl auch große Softwarefirmen dieser Aufgabe annehmen, entweder mit eigener Software oder durch den Aufkauf bestehender Firmen. Auch wird im Moment die Entwicklung meist nur von Informatikern vorangetrieben, die zwar Experten für Softwareentwicklung sind, allerdings kaum tiefergehende Kenntnisse im Bereich menschliche Sprache besitzen. Für die Weiterentwicklung ist es daher sehr wichtig auch Sprachwissenschaftler in die Entwicklung mit einzubeziehen. Erst dann werden wohl auch die Feinheiten der menschlichen Sprache verlässlich abgebildet werden können. Dieser Sachverhalt könnte auch erklären, warum lernbasierte Systeme momentan meist besser abschneiden als semantikbasierte Systeme. Bei lernbasierten Systemen können Informatiker das Erkennungssystem als Blackbox verwenden. Sie müssen nicht verstehen, welche Teile der Sprache zu welchen Zuordnungen führen, da hier nur Ähnlichkeiten gemessen werden. Bei semantikbasierten Systemen ist es dagegen zwingend erforderlich die Bedeutung von Sprachelementen zu kennen.

Ob sich Dinge wie Sarkasmus jemals zuverlässig maschinell erkennen lassen, bleibt weiterhin offen. Sprachliche Elemente wie Sarkasmus beruhen nicht auf speziellen Wörtern oder Satzstellungen, sie leiten sich aus dem Kontextwissen her. Eine Maschine müsste für die Erlangung dieses Kontextwissens über alle aktuellen Gegebenheiten informiert sein, um Aussagen richtig interpretieren zu können. Desweiteren muss ein maschineller Abgleich des gespeicherten Wissens mit dem des zu analysierenden möglich sein, um den Wahrheitsgehalt einer Aussage überprüfen zu können. Spätestens wenn nicht öffentlich verfügbares

Insiderwissen vonnöten ist, um eine Aussage richtig interpretieren zu können, wird ein Algorithmus zwangsläufig scheitern, daraus den Sarkasmus erkennen zu können. Jedoch selbst wenn sich die Erkennung von Sarkasmus als nicht machbar erweisen sollte, verschlechtert sich die Klassifikationsgenauigkeit eines Algorithmus damit insgesamt nur marginal, da nur ein geringer Teil aller Aussagen Sarkasmus enthält.

Im Gegensatz dazu könnte die maschinelle Erkennung von Glaubwürdigkeit durchaus eine Zukunft haben. Carlos Castillo u.a. konnten zeigen, dass die soziale Netzwerkstruktur von Twitter Hinweise auf die Glaubwürdigkeit in Form von Links, Retweets und Nutzereigenschaften bietet vgl. (Castillo, et al., 2011). Derartige Informationen können zwar keine absolut verlässliche Aussage über die Glaubwürdigkeit erbringen, allerdings geben sie eine Tendenz an, die über die Gesamtheit aller Aussagen durchaus nützlich sein kann. Neben der reinen Meinungsrichtung, könnte die Glaubwürdigkeit somit als ein Gewichtungsfaktor verwendet werden.

Wie bereits erwähnt, ist die Einbindung verlinkter Webseiten in der derzeitigen Forschung noch nicht umgesetzt worden. Durch den sehr hohen Anteil an Tweets, deren Interpretation hierdurch erschwert wird, ist dies ein Mangel, der definitiv korrigiert werden muss. Entsprechende Untersuchungen müssen hier allerdings noch die Machbarkeit dieses Vorgehens beweisen. Das Einlesen verlinkter Daten an sich stellt hierbei keine große Schwierigkeit dar, die Interpretierung der Daten allerdings schon. Es wird sich zeigen müssen, mit welcher Genauigkeit man derartige Daten verarbeiten kann. Formatierungsinformationen einer Webseite lassen sich zwar sehr leicht entfernen, der Inhalt dagegen müsste aufgrund seiner Position auf der Webseite und dem darin enthaltenen Text interpretiert werden.

Die wichtigste Verbesserung, die bei der Analyse von Tweets noch offen steht, ist eine bessere Erkennung relevanter Nachrichten. Die manuelle Überprüfung des Datenmaterials hat gezeigt, dass das Konzept der Hashtags in Twitter nicht ausreicht, um relevante Nachrichten eindeutig zu identifizieren. Vor allem populäre Hashtags wie #Google werden, wie bereits erwähnt, sehr häufig nur alles Verbreitungshilfe genutzt. Eine Möglichkeit diesem Umstand entgegenzuwirken wäre eine Liste an Wörtern zu erstellen, die den Bezug zu einem gewünschten Begriff mit hoher Wahrscheinlichkeit herstellen oder ausschließen können. Hierzu müsste allerdings eine sehr große Anzahl an Tweets manuell überprüft und entsprechende Wörter herausgefiltert werden. Da für jeden gewünschten Begriff eine unterschiedliche Menge an Wörtern deren Relevanz angibt, muss für jeden neuen Suchbegriff eine neue Liste an Wörtern erstellt werden. Eine generische Relevanzprüfung ist hierdurch nicht möglich.

Eine positive Entwicklung im Hinblick auf mögliche Analysen ist die Tatsache, dass immer mehr Anbieter sozialer Webdienste anfangen, ihre Daten durch eine API öffentlich zugänglich zu machen. Hierdurch steigt die verfügbare Datenmenge und somit auch die Möglichkeiten für die Analyse dieser Daten.

Abschließend kann man sagen, dass eine Sentimentanalyse auf soziale Medien wie Twitter zwar noch unter ein paar Kinderkrankheiten leidet, allerdings können diese durchaus gelöst werden. Was Twitter anbelangt, hat der Vergleich mit der US Zensus Studie gezeigt, dass

Twitter durchaus Personen aus allen Teilen der Gesellschaft beinhaltet und nicht nur ein Phänomen der Jugend ist. Bei dem weiter steigenden Aufkommen an Tweets werden die Nachrichten in Twitter somit die Realität immer besser abbilden können. Ob eine Analyse aufgrund sozialer Medien-Daten allerdings auch eine reelle Vorhersagekraft auf einen Aktienkurs hat, muss wie bereits erwähnt über einen längeren Zeitraum überprüft werden.

7 Literaturverzeichnis

AG Algorithmisches Lernen, T. K., 2001. *uni-kl.de.* [Online]
Available at: http://www-agrw.informatik.uni-kl.de/damit/b/bayes.html
[Zugriff am 05 03 2012].

Anon., 2011. *Wikipedia.* [Online]
Available at: http://de.wikipedia.org/wiki/Part-of-speech_Tagging
[Zugriff am 06 01 2012].

Berger, A. L., Della Pietra, S. A. & Della Pietra, V. J., 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, Vol. 22*, pp. 39-71.

Beverungen, G. & Kalita, J., 2011. *Evaluating Methods for Summarizing Twitter Posts.* [Online]
Available at: http://www.cs.uccs.edu/~kalita/work/reu/REU2011/FinalPapers/Beverungen.pdf
[Zugriff am 08 02 2012].

Blensberg, F., 2001. *Web Log Mining als Instrument der Marketingforschung . Ein systemgestaltender Ansatz für internetbasierte Märkte.* Wiesbaden: Deutscher Universitäts-Verlag.

Bock, H. H., 1974. *Automatische Klassifikation.* Göttingen: Vandenhoeck & Rubrecht.

Böhringer, M., 2009. *Enterprise Microblogging: Grundlagen, Einsatzpotentiale und Konzepte zur Unterstützung wissensbasierer Projektkommunikation.* s.l.:Volker Derballa.

Bollen, J., Mao, H. & Zeng, X.-J., 2011. Twitter mood predicts the stock market. *Journal of Computational Science, 2(1),* 03, pp. 1-8.

Burkert, G., 1996. *Repräsenation von lexikalisch-semantischem Wissen in einem System zur Verarbeitung natürlicher Sprache.* s.l.:Infix Verlag.

Castillo, C., Mendoza, M. & Poblete, B., 2011. Information credibility on twitter.. *WWWACM*, pp. 675-684.

Catone, J., 2008. *How We Tweet: The Definitive List of the Top Twitter Clients.* [Online]
Available at: http://www.readwriteweb.com/archives/top_twitter_clients_definitive_list.php
[Zugriff am 13 04 2012].

Census, U., 2009. *census.gov.* [Online]
Available at: http://www.census.gov/
[Zugriff am 10 03 2012].

Colston, H. L. & Gibbs, R. W., 2007. *Irony in Language and Thought.* s.l.:Lawrence Erlbaum Assoc Inc.

Damasio, A. R., 1994. *Descartes' error : emotion, reason, and the human brain.* : Putnam Publishing.

De Choudhury, M. et al., 2010. *How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?*. s.l.:In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (May 2010) .

Ding, X., Liu, B. & Yu, P. S., 2008. A Holistic Lexicon-Based Approach to Opinion Mining. *WSDM '08 Proceedings of the international conference on Web search and web data mining*, 11 02.

Eagle, N. & Pentland, A., 2006 . Reality Mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 4 3.

emarketer, 2011. *emarketer*. [Online]
Available at: http://www.emarketer.com/Article.aspx?R=1008615&ds
[Zugriff am 13 04 2012].

Fama, E. F., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance: Volume 25*, 05, pp. 383-417.

Flanagin, A. J. & Metzger, M. J., 2000. Perceptions of Internet information credibility. *Journalism and Mass Communication Quarterly*, pp. 515-540.

Forgas, J. P., 1995. Mood and judgment: The affect infusion model (AIM). *Psychological*, Issue 117, pp. 39-66.

Frawley, W. J., Piatesky-Shapiro, G. & Matheus, C. J., 1992. Knowledge Discovery in Database: An Overview. *AI Magazine*.

Fuhr, N., 2010. *Einführung in Information Retrieval Skriptum zur Vorlesung im SS 10.* [Online]
Available at: http://www.is.informatik.uni-duisburg.de/courses/ir_ss10/folien/skript_1-5.pdf
[Zugriff am 26 01 2012].

Gesellschaft für Informatik, 2009. *uni-hildesheim.de/fgir/*. [Online]
Available at: http://www.unihildesheim.de/fgir/
[Zugriff am 13 03 2009].

Gibbs, R. W., 1986. On the psycholinguistics of sarcasm.. *Journal of Experimental Psychology General 105*, pp. 3-15.

Godbole, N., Srinivasaiah, M. & Skiena, S., 2007. *LargeScale Sentiment Analysis for News and Blogs.* [Online]
Available at: http://www.cs.sunysb.edu/~skiena/lydia/sentiment.pdf
[Zugriff am 23 02 2012].

González-Ibáñez, R., Muresan, S. & Wacholder, N., 2011. Identifying Sarcasm in Twitter: A Closer Look.. *ACL Short Papers The Association for Computer Linguistics*, pp. 581-586.

Gottron, T., 2010. *Information Retrieval.* [Online]
Available at: http://www1.informatik.uni-mainz.de/lehre/ir/skript-sose-10/IR-SoSe10.pdf
[Zugriff am 26 01 2012].

Group Studie Miniwatts Marketing, 2011. *internetworldstats.com.* [Online]
Available at: http://www.internetworldstats.com/stats.htm
[Zugriff am 24 02 2012].

Hajo, H. & Rentzmann, R., 2006. Text Mining. *Informatik Spektrum, Vol. 29(Nr. 4)*, p. 287–290.

Hearst, M., 2003. *berkeley.edu.* [Online]
Available at: http://people.ischool.berkeley.edu/~hearst/text-mining.html
[Zugriff am 2012 01 21].

Heyer, G., 2010. *Text Mining - Wissensrohstoff Text.* [Online]
Available at: http://wortschatz.uni-leipzig.de/~sbordag/TM07/TM01_TextWissenTM.pdf
[Zugriff am 26 01 2012].

Hollmann, F., 2011. *tagesschau.de.* [Online]
Available at: http://www.tagesschau.de/ausland/weibo100.html
[Zugriff am 04 01 2012].

Hong, L., Convertino, G. & Chi, E. H., 2011. *Language Matters in Twitter: A Large Scale Study.* Barcelona, Spanien: International AAAI Conference on Weblogs and Social Media (ICWSM'11).

Java, A., Xiaodan, S., Finin, T. & Tseng, B., 2007. Why we twitter: Understanding microblogging usage and communities. *9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis*, p. 56–65.

Kahneman, D. & Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 03, p. 263–291.

Koski, J. L., Rice, E. M. & Tarhouni, A., 2004. *Noise Trading and Volatility: Evidence from Day Trading and Message Boards.* [Online]
Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=533943
[Zugriff am 2012 01 28].

Kreuz, R. J. & Glucksberg, S., 1989. How to be sarcastic: The echoic reminder theory of verbal irony.. *Journal of Experimental Psychology:*, pp. 374-386.

Kwak, H., Lee, C., Park, H. & Moon, S., 2010. What is Twitter, a social network or a news media?. *WWW '10: Proceedings of the 19th international conference on World wide web*, pp. 591--600.

Leber, J., 2012. *technologyreview.com.* [Online]
Available at: http://www.technologyreview.com/computing/40330/?p1=A1
[Zugriff am 02 05 2012].

Lehner, F. & Maier, R., 1994. *Informationen in Betriebswirtschaftslehre, Informatik und Wirtschaftsinformatik.* Koblenz: Lehrstuhl für Wirtschaftsinformatik und Informationsmanagement, Wiss. Hochsch. für Unternehmensführung.

Lischka, K., 2012. *Roboter machen Ihre Tweets zu Geld.* [Online]
Available at: http://www.spiegel.de/netzwelt/web/0,1518,817082,00.html
[Zugriff am 24 02 2012].

Liu, B., 2009. *Opinion Mining.* s.l.:Springer US.

McCloskey, D. & Arjo, K., 1995. One quarter of GDP is Persuasion. *American*, Issue 85, pp. 191-195.

McGuire, W. J., 1985. Attitudes and attitude change. *Handbook of Social Psychology*, Issue Vol. 2, pp. 233-346.

Meeyoung, C., Haddadi, H., Benevenuto, F. & Gummadi, K. P., 2010. *Measuring user influence in Twitter: The million follower fallacy.* Washington, DC: international AAAI Conference on Weblogs and Social Media (ICWSM).

Mehra, N., Khandelwal, S. & Patel, P., 2002. *Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews.* [Online]
Available at: http://www.stanford.edu/class/archive/cs/cs276a/cs276a.1032/projects/reports/nmehra-kshashi-priyank9.pdf
[Zugriff am 07 03 2012].

Messerschmidt, C. M., Berger, S. C. & Skiera, B., 2010. *Web 2.0 im Retail Banking Einsatzmöglichkeiten Praxsibeispiele und empirische Nutzeranalyse.* Wiesbaden: Gabler Verlag Springer Fachmedien.

Metaxas, P. T., Mustafaraj, E. & Gayo-Avello, D., 2011. *How (Not) To Predict Elections.* [Online]
Available at: http://cs.wellesley.edu/~pmetaxas/How-Not-To-Predict-Elections.pdf
[Zugriff am 17 01 2012].

Miller, M., 2011. *Facebook for Grown-Ups: Use Facebook to Reconnect with Old Friends, Family and Co Workers.* s.l.:Pearson Education Inc..

Nardi, B. A., Schiano, D. J., Gumbrecht, M. & Swartz, L., 2007. *psych.stanford.edu.* [Online]
Available at: http://psych.stanford.edu/~mgumbrec/Why_We_Blog.pdf
[Zugriff am 18 01 2012].

Nofsinger, J., 2005. Social Mood and Financial Economics. *Journal of Behaviour Finance Vol. 6*, pp. 144-160.

Ohana, B. & Tierneyy, B., 2009. *Sentiment classification of reviews using SentiWordNet.* [Online]
Available at: http://arrow.dit.ie/ittpapnin/13/
[Zugriff am 15 02 2012].

Pandey, V. & Iyer, K. C., 2009. *Sentiment Analysis of Microblogs.* [Online]
Available at: http://cs229.stanford.edu/proj2009/PandeyIyer.pdf
[Zugriff am 24 02 2012].

Pang, B., Lee, L. & Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79-86.

PearAnalytics, 2009. *pearanalytics.com.* [Online]
Available at: http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf
[Zugriff am 17 01 2012].

Peterson, H., 2005. *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur.* München: Oldenbourg.

Pew Research Center, f. t. P. &. t. P., 2008. *Internet Overtakes Newspapers as News Outlet.* [Online]
Available at: http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-news-source
[Zugriff am 16 01 2012].

Poblete, B., Garcia, R., Mendoza, M. & Jaimes, A., 2011. Do all birds tweet the same?: characterizing twitter around the world.. *CIKMACM*, pp. 1025-1030.

Princeton Survey Research, A. I., 2005. *Leap of faith: Using the internet despite the dangers.* [Online]
Available at: http://www.consumerwebwatch.org/pdfs/princeton.pdf
[Zugriff am 16 01 2012].

Rettberg, J. W., 2008. *Blogging Digital Media and Social Series.* Cambridge UK: Polity Press.

Richter, M. & Koch, A., 2007. *Enterprise 2.0: Planung, Einführung und erfolgreicher Einsatz von Social Software in Unternehmen.* s.l.:Oldenbourg.

Rittig, J., 2009. *Twitter in der politischen Kommunikation: Analyse von Twitteraktivitäten ausgewählter Politiker während des Landtagswahlkampfes 2009.* s.l.:Grin Verlag.

Sagolla, D., 2009. *140characters.com.* [Online]
Available at: http://www.140characters.com/2009/01/30/how-twitter-was-born/
[Zugriff am 02 01 2012].

Sanders, E. M., 1993. Stock prices and wall street weather. *American Economic Review,* Issue 83, pp. 1337-1345.

Schmitt, I., 2005. *Ähnlichkeitssuche in Multimedia-Datenbanken.* München: Oldenbourg.

Scholz, T., 2011. *Ein Ansatz zu Opinion Mining und Themenverfolgung für eine Medienresonanzanalyse.* [Online]
Available at: http://ceur-ws.org/Vol-733/paper_scholz.pdf
[Zugriff am 01 02 2012].

Schulte im Walde, S., Cramer, I. & Schacht, S., 2004. *uni-saarland.de.* [Online]
Available at: http://www.coli.uni-saarland.de/~schulte/Teaching/Klassifikation-04/bayes-mle.pdf
[Zugriff am 07 03 2012].

Schwarz, G., 2012. *Die Religion des Geldes.* Wien Österreich: Springer Gabel.

Shiller, R. J., 1984. Stock prices and social dynamics,. *Brookings Papers on Economic Activity,* Issue 2, pp. 457-498.

Spiegel Online, 2012. *Deutschland surft amerikanisch.* [Online]
Available at: http://www.spiegel.de/netzwelt/netzpolitik/0,1518,810131,00.html
[Zugriff am 27 01 2012].

Sprenger, T. O. & Welpe, I. O., 2010. *Tweed and Trades, The Information Content of Stock Microblogs.* München: s.n.

Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.-N., 2000. *Web Usage Mining Discovery and Applikations uf Usage Patterns from Web Data,* Minneapolis: Department of Computer Science and Engineering, University of Minnesota.

Takahashi, T., Abe, S. & Igata, N., 2011. Can Twitter Be an Alternative of Real-World Sensors?. *HCII11 Proceedings of the 14th international conference on Human-computer interaction,* pp. 240-249.

Taytal, D. & Komaragiri, S., 2009. Comperative Analysis of the Impact of Blogging and Micro-blogging on Market Performance. *International Journal on Computer Science and Engineering Vol.1(3)*, pp. 176-182.

Tepperman, J., Traum, D. R. & Narayanan, S., 2006. "yeah right": sarcasm recognition for spoken dialogue systems.. *InterSpeech ICSLP*.

Thelwall, M., 2011. *Introduction to Sentiment strength detection.* [Online]
Available at: http://romip.ru/russiras/doc/slides-2011/SentA-russir-day1.pdf
[Zugriff am 17 04 2012].

Tochtermann, K. & Back, A., 2009. *Web 2.0 in der Unternehmenspraxis: Grundlagen, Fallstudien und Trends zum Einsatz von Social Software.* s.l.:Oldenbourg.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M., 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178-185.

Twitter, kein Datum *Twitter.com.* [Online]
Available at: https://support.twitter.com/articles/496007-was-sind-promoted-trends
[Zugriff am 2012 01 30].

Uhr, W., Esswein, W. & Schoop, E., 2003. *Wirtschaftsinformatik.* Dresden: Physica-Verlag HD.

Utsumi, A., 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, pp. 1777-1806.

Vucetic, S., 2003. *dabi.temple.edu.* [Online]
Available at: http://www.dabi.temple.edu/~vucetic/cis526fall2003/lecture1.pdf
[Zugriff am 16 04 2012].

Whitelaw, C., Garg, N. & Argamon, S., 2005. *Using Appraisal Groups for Sentiment Analysis.* [Online]
Available at: http://lingcog.iit.edu/doc/appraisal_sentiment_cikm.pdf
[Zugriff am 08 02 2012].

wikipedia, 2012. *wikipedia.org.* [Online]
Available at: http://de.wikipedia.org/wiki/Maschinelles_Lernen
[Zugriff am 29 02 2012].

Wikipedia, 2012. *wikipedia.org.* [Online]
Available at: http://de.wikipedia.org/wiki/Blog
[Zugriff am 27 01 2012].

Zaïane, O. R., 2001. *Web Usage Mining for a Better Web-Based Learning Environment.* Edmonton, Alberta, Canada: s.n.

Zhang, L. et al., 2011. *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analyse.* [Online]
Available at: http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf
[Zugriff am 08 02 2012].

8 Anhang

```
package TwitterStreamingClient.TwitterStreamingClient;

import java.sql.Timestamp;
import twitter4j.FilterQuery;
import twitter4j.Status;
import twitter4j.StatusAdapter;
import twitter4j.StatusDeletionNotice;
import twitter4j.StatusListener;
import twitter4j.TwitterException;
import twitter4j.TwitterStream;
import twitter4j.TwitterStreamFactory;

public final class PrintSampleStream extends StatusAdapter {

      public static void main(String[] args) throws TwitterException{

            //String[] args2 = new String[1];
            //args2[0] = "#Twitter";
      if(args.length==0){
            System.out.print("Please execute programm with parameters");
            System.exit( 0 );
      }
      final String finalSearchTerm;
      String searchTerm="";
      for(int i=0; i<args.length; i++){
            searchTerm+=args[i]+",";
      }
      finalSearchTerm = searchTerm;


          final Connectdb db = new Connectdb();
            db.connect();


            StatusListener listener = new StatusListener() {
            public void onStatus(Status status) {
               //System.out.println("@" + status.getUser().getScreenName()
+ " - " + status.getText());

                long id=0;
                long user_id=0;
                String user_screen_name="";
                String user_name="";
                String user_description="";
                String user_location="";
                String user_language="";
                Timestamp user_profile_created_at= new
java.sql.Timestamp(status.getUser().getCreatedAt().getTime());
                int user_followers_count=0;
                int user_friends_count=0;
                int user_listted_count=0;
                String text="";
                Timestamp tweet_created_at= new
java.sql.Timestamp(status.getCreatedAt().getTime());
                String profile_image_url="";
                String search_Term = finalSearchTerm;
                Double latitude=0.0;
                Double longitude=0.0;
                String place_country="";
                String place_CountryCode="";
```

```
                String place_FullName="";
                String place_Name="";
                String place_PlaceType="";
                String place_StreetAddress="";
                String in_Reply_To_Screen_Name="";
                long in_Reply_To_Status_Id=0;
                long in_ReplyTo_User_Id=0;
                long re_tweet_Count=0;
                String source="";

                try{  id = status.getId(); } catch (NullPointerException
nulexc) {}
                try{  user_id = status.getUser().getId(); } catch
(NullPointerException nulexc) {}
                try{  user_screen_name = status.getUser().getScreenName(); }
catch (NullPointerException nulexc) {}
                try{  user_name = status.getUser().getName(); } catch
(NullPointerException nulexc) {}
                try{  user_description = status.getUser().getDescription();
} catch (NullPointerException nulexc) {}
                try{  user_location = status.getUser().getLocation(); }
catch (NullPointerException nulexc) {}
                try{  user_language = status.getUser().getLang(); } catch
(NullPointerException nulexc) {}
                //try{  user_profile_created_at = new
java.sql.Timestamp(status.getUser().getCreatedAt().getTime()); } catch
(NullPointerException nulexc) {}
                try{  user_followers_count =
status.getUser().getFollowersCount(); } catch (NullPointerException nulexc)
{}
                try{  user_friends_count =
status.getUser().getFriendsCount(); } catch (NullPointerException nulexc)
{}
                try{  user_listted_count =
status.getUser().getListedCount(); } catch (NullPointerException nulexc) {}
                try{  text = status.getText(); } catch (NullPointerException
nulexc) {}
                //try{  tweet_created_at = new
java.sql.Timestamp(status.getCreatedAt().getTime()); } catch
(NullPointerException nulexc) {}
                try{  profile_image_url =
status.getUser().getProfileImageURL().toString(); } catch
(NullPointerException nulexc) {}
                try{  latitude = status.getGeoLocation().getLatitude(); }
catch (NullPointerException nulexc) {}
                try{  longitude = status.getGeoLocation().getLongitude(); }
catch (NullPointerException nulexc) {}
                try{  place_country = status.getPlace().getCountry(); }
catch (NullPointerException nulexc) {}
                try{  place_CountryCode =
status.getPlace().getCountryCode(); } catch (NullPointerException nulexc)
{}
                try{  place_FullName = status.getPlace().getFullName(); }
catch (NullPointerException nulexc) {}
                try{  place_Name = status.getPlace().getName(); } catch
(NullPointerException nulexc) {}
                try{  place_PlaceType = status.getPlace().getPlaceType(); }
catch (NullPointerException nulexc) {}
                try{  place_StreetAddress =
status.getPlace().getStreetAddress(); } catch (NullPointerException nulexc)
{}
```

```
                try{  in_Reply_To_Screen_Name =
status.getInReplyToScreenName(); } catch (NullPointerException nulexc) {}
                try{  in_Reply_To_Status_Id = status.getInReplyToStatusId();
} catch (NullPointerException nulexc) {}
                try{  in_ReplyTo_User_Id = status.getInReplyToUserId(); }
catch (NullPointerException nulexc) {}
                try{  re_tweet_Count = status.getRetweetCount(); } catch
(NullPointerException nulexc) {}
                try{  source = status.getSource(); } catch
(NullPointerException nulexc) {}


             db.setTweets(
                             id,  //long
                             user_id, //long
                             user_screen_name, //String
                             user_name, //String
                             user_description, //String
                             user_location, //String
                             user_language, //String
                             user_profile_created_at, //Timestamp
                             user_followers_count, //Int
                             user_friends_count, //Int
                             user_listted_count, //Int how many lists a user
belongs to
                             text, //String
                             tweet_created_at, //Timestamp
                             profile_image_url, //String
                             search_Term, //String
                             latitude, //Double
                             longitude, //Double
                             place_country, //String
                             place_CountryCode, //String
                             place_FullName, //String
                             place Name, //String
                             place_PlaceType, //String
                             place_StreetAddress, //String
                             in_Reply_To_Screen_Name, //String
                             in_Reply_To_Status_Id, //long Ist -1 bei null
Werten
                             in_ReplyTo_User_Id , //long Ist -1 bei null
Werten
                             re_tweet_Count,//long
                             source //String
             );
           }

             public void onDeletionNotice(StatusDeletionNotice
statusDeletionNotice) {
                 System.out.println("Got a status deletion notice id:" +
statusDeletionNotice.getStatusId());
             }

             public void onTrackLimitationNotice(int
numberOfLimitedStatuses) {
                 System.out.println("Got track limitation notice:" +
numberOfLimitedStatuses);
             }

             public void onScrubGeo(long userId, long upToStatusId) {
                 System.out.println("Got scrub_geo event userId:" + userId +
" upToStatusId:" + upToStatusId);
```

```
            }

            public void onException(Exception ex) {
                ex.printStackTrace();
            }
        };

        TwitterStream twitterStream = new
TwitterStreamFactory().getInstance();
        twitterStream.addListener(listener);

        FilterQuery fq = new FilterQuery();
        //String[] query = {"#TWitter"};

        fq.track(args);
        twitterStream.filter(fq);

        //twitterStream.sample();
    }
}
```

```
package TwitterStreamingClient.TwitterStreamingClient;


import java.sql.*;


public class Connectdb {

      Connection connection;
    Statement stmt;



void setTweets(
            long MESSAGE_ID,
        long CREATOR_ID,
        String CREATOR_SCREEN_NAME,
        String CREATOR_NAME,
        String CREATOR_DESCRIPTION,
        String CREATOR_LOCATION,
        String user_language,
        Timestamp CREATOR_CREATED_AT,
        int CREATOR_FOLLOWERS_COUNT,
        int CREATOR_FRIENDS_COUNT,
        int CREATOR_LISTTED_COUNT,
        String TEXT,
        Timestamp TWEET_CREATED_AT,
        String CREATOR_IMAGE_URL,
        String SEARCH_TERM,
        Double LATITUDE,
        Double LONGITUDE,
        String PLACE_COUNTRY,
        String PLACE_COUNTRY_CODE,
        String PLACE_FULL_NAME,
        String PLACE_NAME,
        String PLACE_PLACE_TYPE,
        String PLACE_STREET_ADDRESS,
        String IN_REPLY_TO_SCREEN_NAME,
        long IN_REPLY_TO_STATUS_ID,
        long IN_REPLY_TO_USER_ID,
```

```
        long RE_TWEET_COUNT,
        String TWEET_SOURCE
){

      try {
            //System.out.println("Beginn Insert into Tweets:");
            String sql="INSERT INTO MASTER_THESIS_TWEETS (" +
                 "MESSAGE_ID, " +
                       "CREATOR_ID, " +
                       "CREATOR_NAME, " +
                       "CREATOR_SCREEN_NAME, " +
                       "CREATOR_LOCATION, " +
                       "TEXT, " +
                       "SEARCH_TERM, " +
                       "TWEET_SOURCE, " +
                       "PLACE_COUNTRY, " +
                       "PLACE_NAME, " +
                       "CREATOR_DESCRIPTION, " +
                       "CREATOR_CREATED_AT, " +
                       "CREATOR_FOLLOWERS_COUNT, " +
                       "CREATOR_FRIENDS_COUNT, " +
                       "CREATOR_LISTTED_COUNT, " +
                       "TWEET_CREATED_AT, " +
                       "CREATOR_IMAGE_URL, " +
                       "LATITUDE, " +
                       "LONGITUDE, " +
                       "PLACE_COUNTRY_CODE, " +
                       "PLACE_FULL_NAME, " +
                       "PLACE_PLACE_TYPE, " +
                       "PLACE_STREET_ADDRESS, " +
                       "IN_REPLY_TO_SCREEN_NAME, " +
                       "IN_REPLY_TO_STATUS_ID, " +
                       "IN_REPLY_TO_USER_ID, " +
                       "RE_TWEET_COUNT " +

              ") VALUES
(?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?)";

            PreparedStatement ps = connection.prepareStatement(sql);

            if (MESSAGE_ID != 0){ps.setLong(1, MESSAGE_ID);}else
{ps.setNull(1, java.sql.Types.INTEGER);}
            if (CREATOR_ID != 0){ps.setLong(2, CREATOR_ID);}else
{ps.setNull(2, java.sql.Types.INTEGER);}
            if (CREATOR_NAME != ""){ps.setString(3, CREATOR_NAME);}else
{ps.setNull(3, java.sql.Types.VARCHAR);}
            if (CREATOR_SCREEN_NAME != ""){ps.setString(4,
CREATOR_SCREEN_NAME);}else {ps.setNull(4, java.sql.Types.VARCHAR);}
            if (CREATOR_LOCATION != ""){ps.setString(5,
CREATOR_LOCATION);}else {ps.setNull(5, java.sql.Types.VARCHAR);}
            if (TEXT != ""){ps.setString(6, TEXT);}else {ps.setNull(6,
java.sql.Types.VARCHAR);}
            if (SEARCH_TERM != ""){ps.setString(7, SEARCH_TERM);}else
{ps.setNull(7, java.sql.Types.VARCHAR);}
            if (TWEET_SOURCE != ""){ps.setString(8, TWEET_SOURCE);}else
{ps.setNull(8, java.sql.Types.VARCHAR);}
            if (PLACE_COUNTRY != ""){ps.setString(9, PLACE_COUNTRY);}else
{ps.setNull(9, java.sql.Types.VARCHAR);}
            if (PLACE_NAME != ""){ps.setString(10, PLACE_NAME);}else
{ps.setNull(10, java.sql.Types.VARCHAR);}
            if (CREATOR_DESCRIPTION != ""){ps.setString(11,
CREATOR_DESCRIPTION);}else {ps.setNull(11, java.sql.Types.VARCHAR);}
```

```
            ps.setTimestamp(12, CREATOR_CREATED_AT);
            if (CREATOR_FOLLOWERS_COUNT != 0){ps.setLong(13,
CREATOR_FOLLOWERS_COUNT);}else {ps.setNull(13, java.sql.Types.INTEGER);}
            if (CREATOR_FRIENDS_COUNT != 0){ps.setLong(14,
CREATOR_FRIENDS_COUNT);}else {ps.setNull(14, java.sql.Types.INTEGER);}
            if (CREATOR_LISTTED_COUNT != 0){ps.setLong(15,
CREATOR_LISTTED_COUNT);}else {ps.setNull(15, java.sql.Types.INTEGER);}
            ps.setTimestamp(16, TWEET_CREATED_AT);
            if (CREATOR_IMAGE_URL != ""){ps.setString(17,
CREATOR_IMAGE_URL);}else {ps.setNull(17, java.sql.Types.VARCHAR);}
            if (LATITUDE != 0){ps.setDouble(18, LATITUDE);}else
{ps.setNull(18, java.sql.Types.DOUBLE);}
            if (LONGITUDE != 0){ps.setDouble(19, LONGITUDE);}else
{ps.setNull(19, java.sql.Types.DOUBLE);}
            if (PLACE_COUNTRY_CODE != ""){ps.setString(20,
PLACE_COUNTRY_CODE);}else {ps.setNull(20, java.sql.Types.VARCHAR);}
            if (PLACE_FULL_NAME != ""){ps.setString(21,
PLACE_FULL_NAME);}else {ps.setNull(21, java.sql.Types.VARCHAR);}
            if (PLACE_PLACE_TYPE != ""){ps.setString(22,
PLACE_PLACE_TYPE);}else {ps.setNull(22, java.sql.Types.VARCHAR);}
            if (PLACE_STREET_ADDRESS != ""){ps.setString(23,
PLACE_STREET_ADDRESS);}else {ps.setNull(23, java.sql.Types.VARCHAR);}
            if (IN_REPLY_TO_SCREEN_NAME != ""){ps.setString(24,
IN_REPLY_TO_SCREEN_NAME);}else {ps.setNull(24, java.sql.Types.VARCHAR);}
            if (IN_REPLY_TO_STATUS_ID != 0){ps.setLong(25,
IN_REPLY_TO_STATUS_ID);}else {ps.setNull(25, java.sql.Types.INTEGER);}
            if (IN_REPLY_TO_USER_ID != 0){ps.setLong(26,
IN_REPLY_TO_USER_ID);}else {ps.setNull(26, java.sql.Types.INTEGER);}
            if (RE_TWEET_COUNT != 0){ps.setLong(27, RE_TWEET_COUNT);}else
{ps.setNull(27, java.sql.Types.INTEGER);}

           ps.executeUpdate();

           //System.out.println("Insert into Tweets Tabelle:
Abgeschlossen");
      }catch (Exception exc){
            System.out.println(exc);
            connect();
      }
}


public void disconnectdb(){
      try {
      // Wenn ein Fehler auftritt, Fehler ausgeben und versuchen die
Datenbank-Verbindung zu schließen.
      connection.close();
      //System.out.println("Verbindung abgebaut");
    } catch (SQLException sqlexc) {
      System.err.println("Verbindung konnte nicht geschlossen werden.");
    } catch (NullPointerException nulexc) {
      System.err.println("Es wurde keine Verbindung geoeffnet.");
    }
}

public void connect(){
      try {

        connection = DriverManager.getConnection
        ("jdbc:oracle:thin:@//server:1521/orcl", "Username", "Password");

        //System.out.println("Verbindung aufgebaut");
```

```
        stmt = connection.createStatement();


    } catch (Exception exc) {
            System.err.println("Es ist ein Fehler beim Verbinden zur
Datenbank aufgetreten:\n" + exc.getMessage());
        exc.printStackTrace();


      }
}


public Connectdb() {

}

}
```

```
import java.util.ArrayList;
import java.util.List;
import com.cybozu.labs.langdetect.Detector;
import com.cybozu.labs.langdetect.DetectorFactory;
import com.cybozu.labs.langdetect.LangDetectException;


public class LanguageIdentification {

      public static String detectLanguage(String text) throws
LangDetectException{

            Detector detector = DetectorFactory.create();
            detector.append(text);
            String lang = detector.detect();
            return lang;

      }

      public static void main(String args[]){


            System.out.println("Start Programm");

            try {

      //DetectorFactory.loadProfile("/home/valerius/LanguageIdentifikation/
profiles");

      DetectorFactory.loadProfile("C:\\Users\\Pee\\workspace\\LangIdent\\pr
ofiles");
            } catch (LangDetectException e1) {
                  // TODO Auto-generated catch block
                  //e1.printStackTrace();
            }


            Connectdb connectdb = new Connectdb();
            connectdb.connect();

            List<Tweet> list = new ArrayList<Tweet>();
```

```
            System.out.println("Start Download Tweets");
            list = connectdb.getTweetsDb();
            connectdb.disconnectdb();
            System.out.println("Download Tweets Beendet");
            System.out.println(list.size()+": Datensätze Heruntergeladen");

            System.out.println("Starte Spracherkennung");
            for(int i=0; i<list.size(); i++){

      list.get(i).setText(list.get(i).getText().replaceAll("(\\bhttp://[^\\
s]+)","",""));//Links entfernen

      list.get(i).setText(list.get(i).getText().replaceAll("\\B#(\\w*[a-zA-
Z]+\\w*)","",""));//Hashtags entfernen

      list.get(i).setText(list.get(i).getText().replaceAll("(\\B@\\w*[a-zA-
Z]+\\w*)","",""));//Usernamen Entfernen

      list.get(i).setText(list.get(i).getText().replaceAll("\"",""));
                  try {

      list.get(i).setLanguage(detectLanguage(list.get(i).getText()));
                  } catch (LangDetectException e) {
                        // TODO Auto-generated catch block
                        //e.printStackTrace();
                  }
            }
            System.out.println("Spracherkennung Beendet");
            System.out.println("Sende Daten an Datenbank");

            connectdb.setLanguage(list);

            System.out.println("Senden Beendet");

            System.out.println("Programm Ende");


   }

}
```

```
import java.util.ArrayList;
import java.util.List;
import uk.ac.wlv.sentistrength.SentiStrength;

public class SentiScore {

      //public static String[] array = {"text", "hello", "sentidata",
"C:/SentStrength_Data/"};
      public static String[] array = {"text", "hello", "sentidata",
"/home/valerius/SentimentAnalysis/SentStrength_Data/"};
      public static SentiStrength sentiStrengh = new SentiStrength(array);

      public static String getSentiScore(String text){
            String score= sentiStrengh.computeSentimentScores(text);
            return score;
      }

      public static void main(String args[]){
```

```
      Connectdb connectdb = new Connectdb();

      String score=null;
      List<Tweet> list = new ArrayList<Tweet>();

      System.out.println("Start Download Tweets");
      connectdb.connect();
      list = connectdb.getTweetsDb();
      connectdb.disconnectdb();

      System.out.println("Download Tweets Beendet");
    System.out.println(list.size()+": Datensätze Heruntergeladen");

            System.out.println("Starte Sentimenterkennung");
            for(int i=0; i<list.size(); i++){

      list.get(i).setText(list.get(i).getText().replaceAll("(\\bhttp://[^\\
s]+)","",""));//Links entfernen

      list.get(i).setText(list.get(i).getText().replaceAll("#",""));

      list.get(i).setText(list.get(i).getText().replaceAll("(\\B@\\w*[a-zA-
Z]+\\w*)","",""));//Usernamen Entfernen

      list.get(i).setText(list.get(i).getText().replaceAll("\"",""));

                  score = getSentiScore(list.get(i).getText());

                  list.get(i).setGoodScore(score.substring(0,1));
                  list.get(i).setBadScore(score.substring(2,4));
            }
            System.out.println("Sentimenterkennung Beendet");

            System.out.println("Sende Daten an Datenbank");
            connectdb.setSentiScore(list);
            System.out.println("Senden Beendet");

            connectdb.disconnectdb();
            System.out.println("Programm Ende");

      }
}
```

```
import java.sql.*;
import java.util.ArrayList;
import java.util.List;

public class Connectdb {

      Connection connection;
    Statement stmt;



    List<Tweet> getTweetsDb(){
      List<Tweet> list = new ArrayList<Tweet>();

        long tweetId;
            String TEXT;
```

```
      try {
                  ResultSet rs = stmt.executeQuery("SELECT MESSAGE_ID, TEXT
FROM MASTER_THESIS_TRAININGS_DATEN");

                  while ( rs.next() ) {

                    tweetId = rs.getLong("MESSAGE_ID");
                    TEXT = rs.getString("TEXT");

                    Tweet entry = new Tweet();
                  entry.setTweetId(String.valueOf(tweetId));
                  entry.setText(TEXT);
                    list.add(entry);
                  }


      } catch (SQLException e) {
                  // TODO Auto-generated catch block
                  e.printStackTrace();
            }
      return list;
    }

    List<Double> getSentimentScores(){
      List<Double> list = new ArrayList<Double>();

        try {
                  ResultSet rs = stmt.executeQuery("select
sum(good_score+bad_score) as score from master_thesis_trainings_daten where
language_code = ' en ' group by trunc(tweet_created_at,'dd') order by
trunc(tweet_created_at,'dd')");

                  while ( rs.next() ) {

                        list.add(rs.getDouble("score"));
                  }


      } catch (SQLException e) {
                  // TODO Auto-generated catch block
                  e.printStackTrace();
            }
      return list;
    }

    List<Integer> getTweetsPerDay(){
      List<Integer> list = new ArrayList<Integer>();

        try {
                  ResultSet rs = stmt.executeQuery("select count(*) as
Anzahl from master_thesis_trainings_daten  group by
trunc(tweet_created_at,'dd') order by trunc(tweet_created_at,'dd')");

                  while ( rs.next() ) {

                        list.add(rs.getInt("Anzahl"));
                  }


      } catch (SQLException e) {
                  // TODO Auto-generated catch block
                  e.printStackTrace();
```

```
            }
      return list;
    }

    List<GoogleStock> getStockScores(){
      List<GoogleStock> list = new ArrayList<GoogleStock>();

        try {
                  ResultSet rs = stmt.executeQuery("select * from
MASTER_THESIS_GOOGLE_BOERSENK order by datum");

                  while ( rs.next() ) {

                        GoogleStock googleStock = new GoogleStock();
                        googleStock.setDate(rs.getDate("DATUM"));
                        googleStock.setOpen(rs.getDouble("OPEN"));
                        googleStock.setHigh(rs.getDouble("HIGH"));
                        googleStock.setLow(rs.getDouble("LOW"));
                        googleStock.setClose(rs.getDouble("CLOSE"));
                        googleStock.setVolume(rs.getInt("VOLUME"));

                  list.add(googleStock);
                  }

      } catch (SQLException e) {
                  // TODO Auto-generated catch block
                  e.printStackTrace();
            }
      return list;
    }

public void setSentiScore(List<Tweet> list){
      connect();
      long tweetId=0;
      int goodScore=0;
      int badScore=0;

      for(int i=0; i<list.size(); i++){
            tweetId=Long.valueOf(list.get(i).getTweetId());
            goodScore=Integer.valueOf(list.get(i).getGoodScore());
            badScore=Integer.valueOf(list.get(i).getBadScore());


      if(i%40==0){
            disconnectdb();
            connect();
      }

      try {
             PreparedStatement ps;
             String sql="update MASTER_THESIS_TRAININGS_DATEN set
GOOD_SCORE= ' "+goodScore+" ' where MESSAGE_ID="+tweetId;
             ps = connection.prepareStatement(sql);
             ps.executeUpdate();

             sql="update MASTER_THESIS_TRAININGS_DATEN set BAD_SCORE= '
"+badScore+" ' where MESSAGE_ID="+tweetId;
             ps = connection.prepareStatement(sql);
             ps.executeUpdate();


            }catch (Exception exc){
```

```
                  exc.printStackTrace();
            }
      }

      disconnectdb();

}


public void setLanguage(List<Tweet> list){

      String languageCode="";
      long tweetId=0;
      String text="";

      connect();
      for(int i=0; i<list.size(); i++){
            tweetId=Long.valueOf(list.get(i).getTweetId());
            languageCode=list.get(i).getLanguage();
            text=list.get(i).getText();


            if(i%40==0){
                  disconnectdb();
                  connect();
            }

            try {
                  PreparedStatement ps;
                  String sql="update MASTER_THESIS_TRAININGS_DATEN set
language_code= ' "+languageCode+" ' where MESSAGE_ID="+tweetId;
                  ps = connection.prepareStatement(sql);
                  ps.executeUpdate();

                  //sql="update MASTER THESIS TRAININGS DATEN set text= '
"+text+" ' where MESSAGE_ID="+tweetId;
                  // ps = connection.prepareStatement(sql);
                  // ps.executeUpdate();

                  }catch (Exception exc){
                  exc.printStackTrace();
                  }
      }
      disconnectdb();
}

public void disconnectdb(){
      try {
      // Wenn ein Fehler auftritt, Fehler ausgeben und versuchen die
Datenbank-Verbindung zu schließen.
      connection.close();
      //System.out.println("Verbindung abgebaut");
    } catch (SQLException sqlexc) {
      System.err.println("Verbindung konnte nicht geschlossen werden.");
    } catch (NullPointerException nulexc) {
      System.err.println("Es wurde keine Verbindung geoeffnet.");
    }
}

public void connect(){
      try {
```

```
        connection = DriverManager.getConnection
        ("jdbc:oracle:thin:@//localhost:1521/xe", "twitter", "twitter");

        //System.out.println("Verbindung aufgebaut");
        stmt = connection.createStatement();


    } catch (Exception exc) {
            System.err.println("Es ist ein Fehler beim Verbinden zur
Datenbank aufgetreten:\n" + exc.getMessage());
        exc.printStackTrace();


      }
}


public Connectdb() {

}


}
```

```
public class GoogleStock {

      java.sql.Date date;
      double open;
      double high;
      double low;
      double close;
      int volume;

      public java.sql.Date getDate() {
            return date;
      }
      public void setDate(java.sql.Date date) {
            this.date = date;
      }
      public double getOpen() {
            return open;
      }
      public void setOpen(double open) {
            this.open = open;
      }
      public double getHigh() {
            return high;
      }
      public void setHigh(double high) {
            this.high = high;
      }
      public double getLow() {
            return low;
      }
      public void setLow(double low) {
            this.low = low;
      }
      public double getClose() {
            return close;
```

```
	}
	public void setClose(double close) {
		this.close = close;
	}
	public int getVolume() {
		return volume;
	}
	public void setVolume(int volume) {
		this.volume = volume;
	}
}
```

```
import java.util.HashMap;

public class Tweet{


	private HashMap<String,String> tweetInfo = new HashMap<String,String>
();

	public Tweet() {
		tweetInfo = new HashMap<String,String> ();
		tweetInfo.put("tweetId", "");
		tweetInfo.put("text", "");
		tweetInfo.put("time", "");
		tweetInfo.put("language", "");
		tweetInfo.put("goodScore", "");
		tweetInfo.put("badScore", "");
	}

	public void setTweetId(String tweetId) {
		tweetInfo.put("tweetId", tweetId);
    }
	public void setText(String text) {
		tweetInfo.put("text", text);
    }
	public void setTime(String time) {
		tweetInfo.put("time", time);
    }
	public void setLanguage(String language) {
		tweetInfo.put("language", language);
    }
	public void setGoodScore(String goodScore) {
		tweetInfo.put("goodScore", goodScore);
    }
	public void setBadScore(String badScore) {
		tweetInfo.put("badScore", badScore);
    }


	public String getTweetId() {
	  return  tweetInfo.get("tweetId");
    }
	public String getText() {
	  return  tweetInfo.get("text");
    }
	public String getTime() {
	  return  tweetInfo.get("time");
    }
```

```
      public String getLanguage() {
        return  tweetInfo.get("language");
    }
      public String getGoodScore() {
        return  tweetInfo.get("goodScore");
    }
      public String getBadScore() {
        return  tweetInfo.get("badScore");
    }

}
```