Die Herausforderungen und Lösungsansätze bei der Analyse von Microblog-Statusmeldungen im Kontext von Natural Language Processing

Andreas Greiner - info@novo.at
Johannes Kepler Universität, Linz, Austria
Seminar aus Informatik (Crowd Knowledge Extraction) – 351.096
3. Semester Webwissenschaften
14. Februar 2013

Abstract: Beim Natural Language Processing von Statusmeldungen in Microblogs kommt es zu neuen Herausforderungen, da die dort verwendete Sprache jugendlicher und umgangssprachlicher ist. Welche Herausforderungen und Lösungsansätze in diesem Kontext bestehen, werden in dieser Seminararbeit aufgezeigt.

Keywords: Natural Language Processing, Computerlinguistik, Microblogs

1 Inhaltsverzeichnis

2 Kurzfassung

Das Schreiben von Statusmeldungen in Microblogs wie z. B. Twitter ist von einem Randphänomen zu einem Massenphänomen geworden. Dadurch rücken diese auch in das Interesse der Forschung und Wirtschaft. Im Mittelpunkt dieser Arbeit stehen die Herausforderungen und Lösungsansätze aus der Sicht der Domäne Natural Language Processing bei der Verarbeitung von genau diesen Statusmeldungen in Microblogs.
Dort finden sich in der Sprache eine vereinfachte Syntax, jugendsprachliches Vokabular und grafische Stilmittel für den Ausdruck von Gefühlen. Durch diese Besonderhei-

ten kommt es bei der Datenaufbereitung und bei der darauf folgenden Verarbeitung der Daten zu Herausforderungen, die in dieser Arbeit mit exemplarischen Lösungsansätzen beschrieben werden.

3 Einleitung

Die Art der Kommunikation hat sich durch den Erfolg von Social Network Services verändert und hat sowohl auf Unternehmen als auch auf die Gesellschaft Auswirkungen. Das populärste soziale Netzwerk Facebook hat erst vor wenigen Wochen bekanntgegeben, dass über 1 Milliarde Accounts existieren (Facebook Inc., 2012). Im September 2011 verlautbarte Twitter, dass über 100 Millionen aktive User regelmäßig Kurzbotschaften über diesen Dienst senden (Twitter Inc, 2011). Nach aktuellen Angaben waren es im März 2012 dann bereits 140 Millionen aktive User (Statista.de, 2012) . Diese eindrucksvollen Zahlen aus diesen beiden Netzwerken zeigen, dass es sich hier nicht um ein Randphänomen handelt. Mit diesem Wachstum einher geht das zunehmende Interesse aus Wirtschaft und Wissenschaft an der auf diesen Plattformen stattfindenden Kommunikation. Die auf den sozialen Netzwerken auffindbaren Statusmeldungen sind meist informeller Natur und stellen damit eine besondere Herausforderungen für die bereits bekannten Werkzeuge des Natural Language Processings (NLP) dar, denn die dortige Anwendung von Vokabular und Grammatik unterscheidet sich deutlich (Ritter, Clark, & Etzioni, 2011). Dies zeigt beispielhaft auch die Abbildung 1 aus einem realen Beitrag auf Twitter.

Abbildung 1: Statusmeldung in informeller Sprache[1]

Wie dieses Beispiel gut zeigt, handelt es sich dabei um nicht strukturierte Daten, welche für NLP Zwecke aufbereitet werden müssen, um diese in weiterer Folge verarbeiten zu können. Diese Aufbereitung erfolgt durch den Einsatz verschiedener Methoden wie Stemming oder Tokenizing. Durch den Umstand, dass Statusmeldungen meist kurz gehalten und das private Umfeld betreffen, kommen Tipp-, Grammatik- oder Rechtschreibfehler häufiger vor. Ebenso wird die Anzahl der verwendeten Zeichen für eine Statusmeldung ebenfalls eher knapp gehalten.

3.1 Ziele

Im Zuge dieser Seminararbeit werden die unten angeführten Fragestellungen beantwortet, um folgende globale Frage dieser Arbeit zu beantworten:

[1] https://twitter.com/klein_veri/status/273155234782642176 [28.11.2012]

Welche Herausforderungen und Lösungsansätze bestehen bei der Analyse von Statusmeldungen in Microblogs im Kontext von Natural Language Processing?

Im Verlauf dieser Arbeit werden die folgenden untergeordneten Fragestellungen behandelt und beantwortet:

- Welche typischen Aufgaben bestehen bei der Abarbeitung im Bereich NLP?
- Was sind die Besonderheiten und Herausforderungen aus der Sicht der Linguistik bei den Microblog Statusmeldungen?
- Welche grundsätzlichen Möglichkeiten bestehen, um diese NLP Herausforderungen in dieser Domäne zu begegnen?
- Welche konkreten Werkzeuge stehen dafür zur Verfügung und welche Erkenntnisse gibt es aus deren Einsatz?

3.2 Vorgehensweise

Zu Beginn der Ausarbeitung werden die wichtigsten Aufgaben im Bereich NLP erwähnt, welche bei der Analyse von Texten und im Speziellen bei Statusmeldungen im Microblogging Bereich benötigt werden. Im Anschluss daran werden im Kapitel vier die linguistischen Besonderheiten und Herausforderungen bei typischen Statusmeldungen behandelt.

Im darauf folgenden Kapitel fünf werden die Möglichkeiten und Ansätze zur Aufbereitung von Statusmelden vorgestellt. Darauf folgen eine Vorstellung von Werkzeugen aus dieser Domäne und die Erkenntnisse aus deren Einsatz.

Die Arbeit bezieht sich bei den Statusmeldungen vornehmlich auf Twitter, wobei ähnliches natürlich auch mit anderen Statusmeldungen bzw. Meldungen in Microblogs möglich ist. Der Grund für die Konzentration auf Twitter liegt darin, dass durch die Offenheit und Dynamik von Twitter, es eine relative hohe Anzahl an Publikationen gibt, die sich mit diesen Herausforderungen auseinandersetzen, die in dieser Arbeit behandelt werden.

4 Verfahren und Ansätze im Natural Language Processing

In diesem Kapitel wird die Frage beantwortet, welche typischen Verfahren und Ansätze im Bereich des Natural Language Processings bestehen, um die anstehenden Aufgaben lösen zu können. Des Weiteren werden auch verschiedene Ansätze zur Lösung dieser zuvor beschriebenen Aufgaben behandelt.

Natural Language Processing ist ein wichtiger Bestandteil von Business Intelligence zur Extraktion und Transformation von Daten. Es finden sich dort die Methoden des Data Minings, Text Minings und in letzter Zeit verstärkt das Thema Opinion Mining wieder. Data Mining konzentriert sich auf die Analyse von strukturierten Daten und versucht durch Extraktion und Entdeckung neue Erkenntnissen aus impliziten Daten zu gewinnen (Frawley, Piatetsky-shapiro, & Matheus, 1992). Text Mining hingegen setzt sich mit der Extraktion und Entdeckung von Daten und Informationen

aus unstrukturierten Daten auseinander. Opinion Mining (auch gerne als Synonym für Sentiment Mining verwendet) versucht aus unstrukturierten Daten das Sentiment festzustellen. Das bedeutet, dass mit den unterschiedlichen Verfahren des Opinion Minings die Meinung oder Gefühlslage festgestellt werden soll.

Je nach Ziel werden für diese Methoden unterschiedliche Verfahren verwendet, um das Ziel zu erreichen. In dieser Arbeit werden jene Verfahren genauer untersucht die sich im Bereich Text- und Opinion Mining wiederfinden mit dem Fokus auf unstrukturierte Daten aus Statusmeldungen wie z. B. auf Twitter. Diese Verfahren setzen sich mit der Morphologie („Wortgrammatik"), also der Analyse von Flexionsformen, Wortarten und Wortbildung auseinander. Andere Verfahren dienen zur Erkennung der Syntax (Satzgrammatik) und andere wiederum mit der Semantik. Die nachfolgenden Kapitel 4.1 bis 4.3 wurden nach den Ausführungen von Kalisch erstellt (Kalisch, 2012).

4.1 Morphologische Analyse

Tokenization: Dieses Verfahren stellt fest, welche Zeichen die einzelnen Wörter trennen und liefert als Ergebnis die einzelnen Wörter zurück. Typische Zeichen die Wörter trennen sind das Leerzeichen, Punkte, Gedankenstriche, usw.

Stemming/Lemmatizing: Nach diesem Vorgang ist der nächste Schritt das Stemming bzw. auch Lemmatisierung genannt. Hier werden die Wörter auf den Wortstamm reduziert.

4.2 Syntaktische Analyse

Part-of-Speech Tagging: Beim Part-of-Speech (POS) Tagging werden die unterschiedlichen Wortformen zu den einzelnen Wörtern aus den untersuchten Sätzen zugeteilt. Es wird hier festgestellt, welches Wort das Objekt, Subjekt, Adverb, etc. ist. Das Verfahren wird Annotation von Texten genannt. Ein ähnliche Vorgehensweise wie das POS-Tagging sind N-Gramme, eine statistische Analyse von Texten.

Phrase Recognition: In diesem Verfahren wird festgestellt, wie die einzelnen Wörter in einem Satz zusammengehören. Ergebnis könnte die Erkennung von Nominal- oder Verbphrasen sein (z. B. „Der alte Mann gab seine Schuhe zurück" → es handelt sich um eine Nominalphrase und der Wortstamm „alt" bezieht sich hier auf Mann).

Parsing: Das Parsing ist eine Weiterentwicklung des POS Taggings. Hierbei werden die einzelnen Wörter auf die jeweilige Stellung im Satz zugeordnet.

4.3 Semantische Analyse

Die semantische Analyse hilft dabei festzustellen welche Bedeutung bzw. in welchem Kontext ein Wort verwendet wird. In der deutschen Sprache ist das Wort „Bank“ ein klassisches Beispiel. Das einzelne Wort alleine genügt auf Grund seiner mehrfachen Bedeutung nicht, sondern der Kontext des Wortes verrät mehr über die jeweilige Bedeutung des Wortes. Um die Deutungshoheit zu gewinnen ist eine semantische Analyse notwendig.

4.4 Ansätze und Ergebnis der Analysen

Die Ansätze mit denen diese Analysen durchgeführt werden, können auf Basis

- von wissens- bzw. regelbasierten Ansätzen, oder anhand
- von Mustersuchen,
- neuronalen Netzen oder auch
- auf Basis von statistischen/korpus-linguistischen Ansätzen

geschehen (Heyer Gerhard, 2011). Eine Kombination der verschiedenen Ansätze zur Verbesserung des Ergebnisses ist natürlich ebenso möglich.

5 Die linguistischen Besonderheiten und Herausforderungen

Die Ziele dieses Kapitels sind das Aufzeigen der Besonderheiten und der Herausforderungen bei Statusmeldungen in Microblogs anhand des Beispiels von Tweets in Twitter.

5.1 Allgemeine Verwendungsmöglichkeiten von Twitter

Die Länge der Tweets ist bei Twitter auf maximal 140 Zeichen begrenzt. Die meisten Tweets haben sogar eine geringere Länge (Go, Bhayani, & Huang, 2009). Twitter selbst bietet verschiedene Features an, welche in einem Tweet vorkommen können und so Bestandteil des Satzes bzw. der Botschaft werden, aber für die Analyse nicht oder bedingt notwendig sind und deshalb gefiltert werden sollten (Owoputi, O’Connor, & Dyer, 2012). Die gängigsten Features lassen sich wie folgt beschreiben:

- Öffentliche Antworten (Reply @user): Hierbei handelt es sich um eine öffentliche Antwort an einen User auf einen Tweet.
- Kopieren und weiterverbreiten (RT @user): Das Aufgreifen eines fremden Tweets und das folgende Weiterverbreiten des Tweets wird als Re-Tweeting bezeichnet und wird mit „RT“ abgekürzt.
- Kennzeichnen und kommentieren mittels Hashtag (#hashtag): Der Hashtag ist im Gegensatz zu den beiden vorangegangenen Funktionen nicht in Twitter selbst inkludiert, sondern ist eine Konvention, welche zum Klassifizieren eines Beitrags verwendet wird (Institut für Technikfolgen-Abschätzung der österreichischen Akademie der Wissenschaften, 2009). Dieser Hashtag wird in den Twitterergebnis-

sen mit einem Link hinterlegt und ein Klick darauf, würde Beiträge anführen, welche ebenfalls ihren Tweet mit diesem Hashtag kategorisiert haben.
- URLs: Gerne werden auch URLs in einem Tweet angeführt. Diese werden meist über URL-Shortener Dienste wie bit.ly nochmals kaskadiert.

5.2 Allgemeines zur verwendeten Sprache in Twitter

Die beiden Medienlinguistiker Peter Koch und Wulf Oesterreicher haben ein für den deutschsprachigen Raum populäres Medienmodell entwickelt. Dieses Schema unterscheidet die verschiedenen Konzepte der Linguistik. Auf der einen Seite, ob die Konzeption eher mündlich orientiert ist (Ausdruck von Nähe) oder doch schriftlich (Ausdruck von Distanz) und auf der anderen Seite ob das Übertragene phonisch oder grafisch ist. Im Falle von Twitter kommt Moraldo zum Entschluss, dass die Übertragung auf Grund des Textes natürlich grafisch ist, es jedoch beim Konzept auf den konkreten Tweet ankommt, denn diese können sowohl mündlich orientiert sein (und daher Nähe vermitteln), aber auch Distanz und folgt somit eher dem Konzept der Schriftlichkeit (Moraldo, 2006). Moraldo hat in seinem Paper *<<das Leben in 140 Zeichen ...heisst Twitter :-)>>* beispielhafte Besonderheiten von Tweets im Gegensatz zur Standardschriftform angeführt, welche sich bei Tweets im privaten Bereich feststellen lassen. Diese Tweets lassen sich nach dem Koch/Oesterreicher Schema bei der mündlichen Konzeption einordnen und ähneln daher der gesprochenen Sprache. Was ist nun so anders bei der verwendeten Sprache in Microblogs?

Vereinfachte Syntax: Bei vielen Tweets fällt es auf, dass es sich selten um einen ganzen vollständigen Satz handelt, sondern gerne verkürzt werden. Dabei werden beispielsweise der Artikel oder das Subjekt weggelassen.

> *"BZÖ schaffts nicht, piraten schon. KPÖ derweil auf platz 2 (!) bei ca. 20 %, verluste für SPÖ, ÖVP und grüne."*[2]

Der Schreibstil ähnelt dem eines Telegrammes. Es wird versucht durch Weglassen von Wörtern Platz zu sparen, damit der Platz von 140 Zeichen optimal ausgenutzt wird. Diese **syntaktischen Reduktionen** sind aber nicht die einzigen, denn es finden sich auch zahlreiche **lexikalische Reduktionen**. Dabei handelt es sich um Abkürzungen die öfters auch aus dem englischsprachigen Raum kommen. So schreibt User @absinthium_ an den User @nutellagangbang in Abbildung 2 unter anderem ROFL (rolling on the floor laughing) und LOL (laughing out loud).

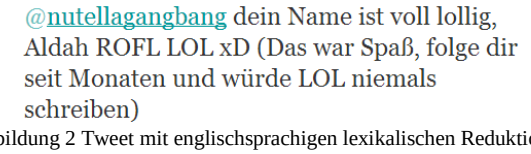

Abbildung 2 Tweet mit englischsprachigen lexikalischen Reduktionen[3]

[2] https://twitter.com/WernerReisinger/status/272737559090700288 [25.11.2012]

Jugendsprachliches Vokabular: Stark beeinflusst vom jungen Alter der typischen Twitter NutzerInnen (Rainie, Brenner, & Purcell, 2012; Statista.de & comScore, 2010) wird natürlich auch die verwendete Sprache von diesen wesentlich beeinflusst. Das jugendsprachliche Vokabular kommt hier zur Anwendung und besticht umgangssprachliche, flotte, stereotype Wörter und Floskeln. Dazu kommen auch diverse Tilgungen von Buchstaben. Dabei wird wie im folgenden Tweet bei dem Wort „nicht" das –t weggelassen oder bei dem Wort „gehe" das letzte –e (Moraldo, 2006).

> *„tasg der offenen tür im schlaflabor. tagespass 15.-€. nich schlecht. vielleicht geh ich hin"*[4]

So verwendet die Userin im Tweet aus Abbildung 3 jugendliches, umgangssprachliches Vokabular mit teilweiser Dialektform wie z. B. „Oida" oder einer e-Tilgung bei „zuck".

> Oida wenn da Papa meine Jacke immer nach
> unten hängt zuck i noch iwann aus ich hab
> dann immer einen stress mitn suchen -.-

Abbildung 3 Jugendliches flappsiger Tweet mit Umgangsspache.[5]

Je nach Region gibt es auch einen stärken Einschlag der Mundart. So informierte die österreichische Mobilfunkmarke TeleRing in einer Presseaussendung über eine Studie, dass im Westen Österreichs Jugendliche ihre SMS eher in Mundart verfassen, als Jugendliche im Osten Österreichs (im Westen sind es 61 % im Gegensatz zum Osten mit 31 %) (TeleRing, 2011). Dazu gibt es natürlich auch die unterschiedlichsten Ausprägungen der Dialekte, wie eine Studie aus den USA beweist (Connor, Smith, & Xing, 2008). Diese zeigt dass im Norden Kaliforniens das Wort „cool" gerne mit „koo" abgekürzt wird, während im Süden von „coo" geschrieben wird. Dazu kommen auch jugendsprachliche Abkürzungen wie „wassup" für „what's up" oder „suttin" für „something".

Grafische Stilmittel: Morallo weist in seinem Paper daraufhin, dass grafische Stilmittel in der computer- und mobilvermittelten Kommunikation eine wichtige Funktion einnehmen zur funktionalen Realisierung der Sprache in diesem Medium. Es handelt sich dabei um eine persönliche, informelle Handschrift des Twitterers (Moraldo, 2006). Bei der Wahl der grafischen Stilmittel kann auf eine Vielzahl an unterschiedlichen Möglichkeiten und Varianten zugegriffen werden, wie

- Smileys: Damit werden Stimmungen und Gefühlszustände in grafischer Form vermittelt. Dabei sind die Smileys bzw. Emoticons um 90° nach rechts zu drehen. Manches Mal werden auch Piktogramme in Textform abgebildet, wie z. B. das Herz mit <3

---

[3] https://twitter.com/absinthium_/status/273028536460263425 [25.11.2012]
[4] https://twitter.com/TheGurkenkaiser/status/273385894927536128 [27.11.2012]
[5] https://twitter.com/Miss_Microsoft/status/273309176699166720 [27.11.2012]

- Emoji: Ein Emoji ist eine japanische Variante eines Smileys (-.- oder x_x).
- Piktogramme: Piktogramme lassen sich mit ALT+eine beliebige Zahl des Nummernblocks erzeugen und sind wiederum eine Möglichkeit Stilmittel in grafischer Form einzusetzen. Beispiele dafür sind: ☺ ☻♥•○
- Mehrfach Iteration: Durch das Wiederholen von Buchstaben oder Buchstabenfolgen werden Wörter hervorgehoben (Beispiele: liiiiiiiiebe, muahahahaha)
- Großschreibung: Durch konsequentes Großschreiben wird die Wichtigkeit hervorgehoben.
- Inflektive: Wortstämme die grundsätzlich als Prädikat verwendet werden, erhalten zu Wortbeginn und –ende einen „*“ um ein Ereignis bildlich zu reproduzieren (z. B. *träum*, *duck*).
- Onomatopoetika: Hier werden charakteristische Verhaltensformen, Bewegungen oder Geräusche in Form eines Tweets nachgeahmt. Dies können z. B. Erstaunen, Lachen („hihihihihi“), Verachtung („ bäh“) oder andere Formen sein.
- Alphanumerische Schreibweisen: Wortbestandteile oder ganze Wörter werden durch die Nutzung von Zahlen ersetzt (z. B. 4ever, Gute N8, 8ung).

Zusammenfassung

Durch die beschränkte Zeichenzahl von max. 140 bei einem Tweet verwenden Twitterer verschiedene Möglichkeiten um diese Tweets prägnanter werden zu lassen. Dabei behelfen Sie sich einer vereinfachten Syntax durch lexikalische und syntaktische Reduktionen. Ebenso zeigt sich auch ein jugendsprachliches Vokabular, gepaart mit regionalen Dialekten und der Verwendung von grafischen Stilmitteln um persönlicher zu wirken und Emotionen zu vermitteln. Auch wenn die Herausforderungen in den verschiedenen Sprachen sich ähneln, so könnte es bei der konkreten Umsetzung zu größeren Herausforderungen kommen durch die nationalen und regionalen Eigenheiten der dort auffindbaren Sprache.

6 Möglichkeiten und Ansätze der Aufbereitung von Statusmeldungen

Dieses Kapitel liefert Antworten auf die Fragestellung, welche grundsätzlichen Möglichkeiten denn bestehen, um die im Kapitel zuvor beschriebenen linguistischen Besonderheiten und Herausforderungen zu bewältigen.

In der Literatur finden sich die hier angeführten Möglichkeiten und Ansätze in der Regel stets im Zusammenhang mit der Bestimmung des Sentiments und wird daher lediglich als Teil des Ganzen gesehen. Der Vorverarbeitung der Daten für das Sentimentsmining kommt jedoch eine wichtige Bedeutung zu. Dieses Kapitel führt die dort angewandten Ansätze an, lässt aber die Schritte zur Bestimmung des Sentiments außen vor, da dies nicht zur Zielerreichung dieser Arbeit beitragen würde.

6.1 Vorverarbeitung der Daten

Im Wesentlichen geht es bei der Vorverarbeitung darum, dass die vorliegenden Daten korrigiert und standardisiert werden zur weiteren Verarbeitung. Wie im Kapitel 5 aufgezeigt bestehen einige linguistische Herausforderungen die es zu bewältigen gibt, denn die Wörter und Sätze welche in Microblogs auffindbar sind, entsprechen meist nicht den gängigen Regeln der Sprache. Die vorgestellten Ansätze beziehen sich auf verschiedene Literaturquellen (Go et al., 2009; Maynard, Bontcheva, & Rout, 2012; Pak & Paroubek, 2010; Ritter et al., 2011; Ritter, Etzioni, & Clark, 2012).

Spracherkennung: Einige Aufgaben in der Vorverarbeitung variieren je nach Sprache. Beispielsweise sind bei der Fehlerkorrektur auf Basis von Rechtschreibfehlern für die deutsche Sprache natürlich andere Regeln gültig, als es für die englische Sprache gilt. Daher gilt es in der Vorverarbeitung festzustellen, welche Sprache im zu untersuchenden Dokument/Korpus vorhanden ist und diese dementsprechend aufzubereiten (Lui & Baldwin, 2011).

Fehlerkorrektur: Das Korrigieren der Fehler mit der Hilfe eines Lexikons hilft Rechtschreib- und Grammatikfehler zu korrigieren. Ebenso werden Abkürzungen und Akronyme ersetzt. Bei diesem Vorgang ist es wichtig, dass auf gute und umfangreiche Rechtschreib- und Grammatikkorrekturen zurückgegriffen werden kann und natürlich auch auf Wörterbücher bzw. Lexika, welche die Langform der Abkürzungen und Akronyme liefert (Kouloumpis, Wilson, & Moore, 2011; Liu, Zhang, Wei, & Zhou, 2011).

Filtern: Das Filtern löscht überflüssige Inhalte. Als überflüssig werden gerne Stop-Wörter oder Punktationen gesehen. Die Gründe hierfür liegen darin, dass bei der Bestimmung des Sentiments, diese Zeichen meist keinen Mehrwert liefern. Wenn die Daten für andere Zwecke verwendet werden würden, so ist je nach Verwendungszweck zu entscheiden, ob das Filtern von bestimmten Zeichen einen Sinn ergibt oder nicht. Beim Filtern der Wörter sollte auch das Löschen von weiteren überflüssigen Zeichen berücksichtigt werden, wie es bei der Mehrfach-Iteration vorkommt. Aus „Scheeeeeeeeeeibe“ wird „Scheibe“. Ein Filter über die typischen Twitter-Features wie Username, Hashtag, Shortener-URL ergibt in den meisten Fällen ebenfalls Sinn, denn speziell bei der Bestimmung des Sentiments führen genau diese Daten zu einem größeren Rauschen. In einem Experiment führte das Filtern von Usernamen, Links und der Mehrfach-Iteration zu einer Reduktion der Datenmenge um über 50 % (Go et al., 2009). Als Technik zur Filterung wird in der Literatur sehr häufig die Verwendung von regulären Ausdrücken (Regex) genannt.

Stemming: Das Stemming führt dazu, dass die vorhandenen verschiedenen Formen eines Wortes auf den Wortstamm reduziert werden mit dem Ziel, dass die verschiedenen Varianten vereinheitlicht werden. Genauso wie bei den Fehlerkorrekturen bedarf

es beim Stemming von Wörtern eines sprachenspezifischen Stemmers (Jiang, Yu, & Zhou, 2011).

6.2 Merkmalsklassifizierung der Daten

Bei der Merkmalsklassifizierung der Daten werden die Eigenschaften der vorliegenden Daten untersucht und festgestellt mit dem Ziel, dass die vorliegenden Daten möglichst genau beschrieben werden.

N-Gramme: Wenn Texte in Fragmente zerlegt werden, so wird in der Linguistik, das Ergebnis als N-Gramme bezeichnet. Das Ergebnis ist ein Wert in Prozenten angegeben, wie häufig ein Buchstabe, Wort bzw. eine Wortkombination in einem existierenden Korpus sich wiederfindet. Das Sentimentsmining geht davon aus, dass Wörter bzw. Wortkombination die einen niedrigen statistischen Wert besitzen für die Feststellung des Sentiments wertvoller sind als jene Wörter die über einen höheren Wert verfügen. Google bietet einen kostenlosen Ngram Viewer an (Michel et al., 2011), welcher grafisch zeigt, wie oft frei eingebbare Wörter im jeweiligen Text-Korpus wiederzufinden ist (in Abbildung 4 wurde ein Vergleich der Wörter „Coca Cola, Fanta und Red Bull“ im deutschen Korpus für die Jahre 1900 bis 2008 durchgeführt).

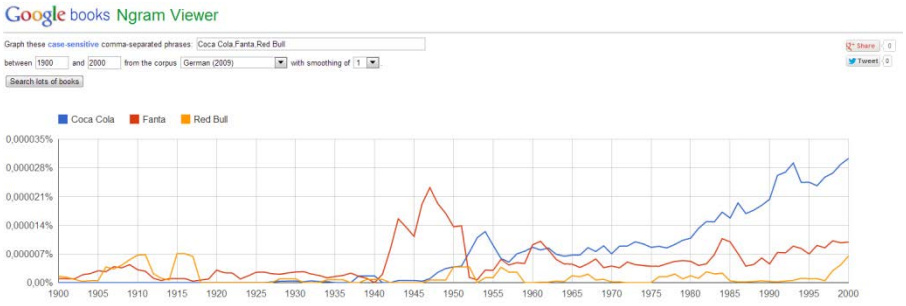

Abbildung 4: N-Gramme Ergebnis von Coca Cola vs. Fanta vs. Red Bull im deutschsprachigem Ngram Google Korpus für die Jahre 1900 - 2008

Bei den N-Grammen wird unterschieden in Uni-Grammen, Bi-Grammen, je nachdem wie viele Buchstaben bzw. Wörter herangezogen werden für die weitere Verarbeitung. Was denn nun besser ist, ob Uni-Gramme, Bi-Gramme, Tri-Gramme, etc. darüber streiten sich die Geister, denn die Forschungsergebnisse sind in dieser Domäne absolut heterogen. Hauptsächlich werden Uni- und Bi-Gramme verwendet. (Go et al., 2009) beispielsweise setzt eine Kombination aus Uni- und Bi-Grammen ein und hatte damit bessere Ergebnisse als Uni- oder Bi-Gramme separat. (Pang, Lee, & Vaithyanathan, 2002) hat einige Jahre zuvor mit Uni-Grammen bessere Ergebnisse erzielen können. (Pak & Paroubek, 2010) führen aus, dass sie mit Bi-Grammen er-

folgreicher waren als mit Uni-Grammen. Der Erfolg richtete sich jeweils auf das Ziel der Bestimmung des Sentiments aus.

TF-IDF: Auch im Bereich Statistik ist tf-idf (term frequency-inverse document frequency) anzutreffen. Das Ergebnis der Berechnung ist eine statistische Zahl, welche die Wichtigkeit eines Worts über ein Dokumentenset angibt. Groot gibt an (Groot, 2012), dass Tf-idf für unsere Zwecke nicht geeignet ist, denn es bekommen bei Tf-idf jene Worte einen hohen Wert die innerhalb eines Tweets oft vorkommen, aber im vorhandenen Korpus selten auftreten. Groot führt hier aus, dass sich diese Wörter nicht dafür eignen, weil diese nur einen kleinen Teil aller Tweets im Korpus widerspiegeln. Der Autor hingegen sieht hier einen ähnlichen Ansatz zu den N-Grammen, wenn auch hier die Term-Frequency innerhalb eines Dokuments deutlich stärker gewichtet wird und somit stark auf das Ergebnis drückt. Was Groot bestätigt ist der Umstand, dass Tf-idf bei anderen Autoren in diesem Bereich ebenfalls keine Rolle spielt.

Part-of-Speech (POS):

Ein Part-of-Speech Tagger ordnet Wortarten zu Wörtern und Satzzeichen zu und berücksichtigt hierbei auch den Kontext. Das Ergebnis eines POS-Taggers ist die „fertige“ Annotation eines Textes. Jedem Wort wird eine bestimmte Wortartkategorie zugeteilt. Im deutschsprachigen Raum führt das Institut für deutsche Sprache in Mannheim zwei aktuelle Tagsets mit den dementsprechenden Taggern. Hierbei handelt es sich um das Tagset Connexor und STTS mit den jeweiligen Taggern CONNEXOR und TreeTagger (Institut für deutsche Sprache, 2012a). Wird ein Text beispielsweise mit dem STTS (Stuttgart Tübingen Tagset) angewendet, so wird kann der untersuchte Text auf drei Kategorien analysiert werden (POS – Wortklasse, LEX – lexikalische Kategorie und MOR – morphologische Merkmale) (Institut für deutsche Sprache, 2012b). Die Abbildung 5 zeigt einen Ausschnitt aus den potentiellen Möglichkeiten wie ein Text annotiert werden könnte.

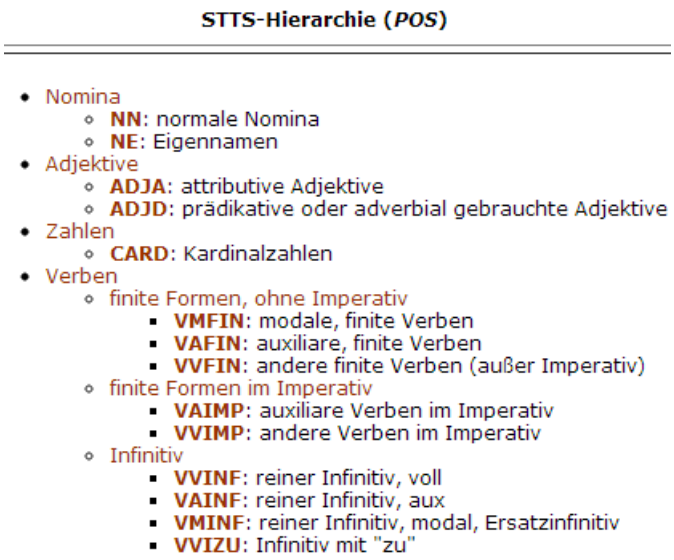

Abbildung 5 Auszug aus möglichen Annotationen im STTS-Tagset (Institut für deutsche Sprache, 2012b)

Probleme die bei einem POS-Tagger auftauchen sind, dass dieser nur für eine Sprache anwendbar ist und dass ein Wort in mehreren Wortartenkategorien vorkommen kann und eine Zuordnung falsch verläuft. Zusätzlich wird ein POS-Tagger auf Basis eines bestimmten Textkorpus trainiert und wenn die zu untersuchenden Dokumente aus einer fremden Domäne stammen, so könnten sich hieraus Herausforderungen ergeben. In dem Paper *"Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics"* erörtert Manning, dass die aktuellen POS-Tagger eine Zuordnungsrichtigkeit von 97 % auf Zeichenebene schaffen. Sieht man sich die Fehlerquote auf Satzebene an, so steigt diese auf 43-45 %. Manning hat zwar durch Verfeinerungen von bestehenden POS-Taggern verbesserte Ergebnisse erhalten (Manning, 2011), jedoch halten sich diese Verbesserungen in Grenzen aus der Sicht des Autors.

POS-Tagging wird von anderen Autoren auch zusätzlich zu den diversen N-Gramm-Ansätzen zusätzlich verwendet, wobei die Ergebnisse im Bereich Sentimentsfeststellung absolut unterschiedlich sind. So schreibt Go et al von schlechteren Ergebnissen (Go et al., 2009), Bethard et al sowie Kouloumpis et al von gleichbleibenden Erkenntnissen beim zusätzlichen Einsatz eines POS-Taggers zur Bestimmung des Sentiments eines Tweets (Bethard, Yu, & Thornton, 2006; Kouloumpis et al., 2011). Dass sich der Einsatz eines POS-Taggers positiv auf die Ergebnisse auswirkt bezeugen die Papers von Pak et al oder Agarwal et al (McCombs & Shaw, 1972; Pak & Paroubek, 2010).

6.3 Zusammenfassung

Das Ergebnis dieses Kapitels ist, dass bei der Vorverarbeitung der Daten immer wieder die gleichen Schritte in den wissenschaftlichen Artikeln zu finden sind. Es handelt

sich dabei um die Erkennung der Sprache des Dokuments, der darauf folgenden Korrektur auf der Basis von gültigen Rechtschreib- und Grammatikregeln sowie der vollständigen Ausschreibung von Abkürzung und Akronymen. Das Filtern ist ebenfalls ein Schritt, der bei der Datenaufbereitung stets durchgeführt wird. Dabei geht es um die Eliminierung der Mehrfach-Iteration von Buchstaben, dem Löschen bzw. Ersetzen von Stopwörtern, Usernamen, Hashtags und Links wie es bei Statusmeldungen in Microblogs üblich ist. Manche Autoren führen auch das Stemming an, um ein dekliniertes Wort auf die Ursprungsform zurückzusetzen.

Wenn die Daten soweit aufbereitet sind, folgt der nächste Schritt, der der Merkmalsklassifizierung. Im Wesentlichen werden hier N-Gramme und das POS-Tagging eingesetzt, wobei die konkrete Vorgehensweise sich hier je nach Autor unterscheidet. Manche Autoren setzen auf Kombinationen bei den N-Grammen (Uni-Gramme, Bi-Gramme, …) und andere auf POS-Tagging bzw. sogar eine Kombination aus der Vorgehensweise N-Gramme gepaart mit POS-Tagging findet sich in der Literatur wieder.

In der Abbildung 6 werden die notwendigen Schritte nochmals in Form einer Abbildung beschrieben, wobei der Schritt der Sentimentsfestellung aus der Behandlung in dieser Arbeit ausgeschlossen wurde und die Konzentration auf die beiden vorgelagerten Prozessschritte gelegt wurde.

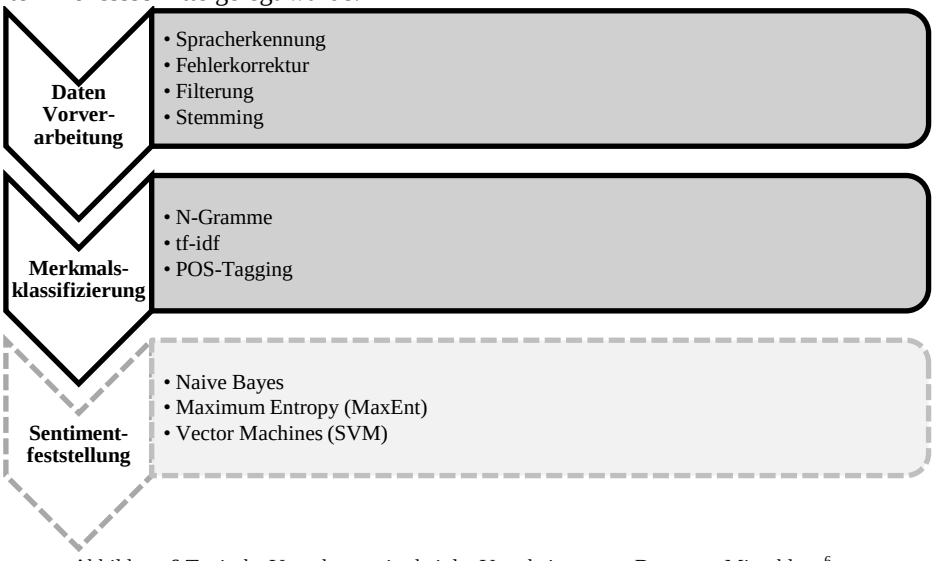

Abbildung 6 Typische Vorgehensweise bei der Verarbeitung von Daten aus Microblogs[6]

7 Werkzeuge und die Erkenntnisse aus deren Anwendung

Diese Kapitel zeigt auf, welche konkreten Werkzeuge und Quellen bei den einzelnen Prozessschritten, welche im vorangegangenen Kapitel 6 erörtert wurden, nun tatsächlich eingesetzt wurden und welche Ergebnisse damit erzielt wurden.

[6] Eigene Abbildung.

Der Hinweis im Artikel von Go et al, dass sie ihre Systeme modular aufbauen, um damit die Möglichkeit zu schaffen, schnell und einfach, verschiedene Module auszutauschen und festzustellen, wie sich deren Adaption auf die Ergebnisse auswirken ist verfolgenswert (Go et al., 2009). Beim Einstieg in diese hier behandelte Domäne, sollte dies ebenfalls auf eine ähnliche Art und Weise über den Aufbau eines modularen Frameworks geschehen. In vielen anderen wissenschaftlichen Artikeln hatte der Autor den Eindruck, dass diese ihre Systeme eher straight-forward aufbauen und sich somit eine gewisse Inflexibilität einschleicht.

7.1 Werkzeuge und Quellen für die Daten-Vorverarbeitung

Spracherkennung: Oft wird bei den wissenschaftlichen Artikeln darauf verwiesen, dass bei der Verwendung der Twitter-API eine Sprachauswahl eingestellt wird. Das bedeutet, dass die Spracherkennung an die Twitter API ausgelagert wird. Was aber, wenn dieser Vorgang selbst gemacht werden möchte? Es stehen dazu beispielhaft folgende Werkzeuge zur Verfügung:

- TextCat (69 Sprachen) - http://odur.let.rug.nl/~vannoord/TextCat/index.html
- LangID (85 Sprachen) - http://langid.net/
- Guesser (~ 100 Sprachen) - http://www.mnogosearch.org/download.html
- Google Translate API (> 100 Sprachen) - https://developers.google.com/translate/

Fehlerkorrektur: Obwohl in den Artikeln öfters angeführt, dass die Korrektur in Form von Rechtschreib- und Grammatikfehlern wichtig ist, so findet sich darin keine Anwendung, die über eine Schnittstelle zu einer automatischen Korrektur der Fehler verfügt.

Filterung: Bei der Filterung setzen alle Autoren auf einen nicht näher spezifizierten Ansatz des Ersetzens der Zeichen. Die Autoren sprechen hier von der Anwendung von regulären Ausdrücken. Die folgenden Autoren haben hier laut den jeweiligen Artikeln folgende Filtermöglichkeiten bei ihren Experimenten eingesetzt (Tabelle 1):

	Go et al, 2009	Ritter et al, 2011	Pak et al, 2010	Kouloumpis et al, 2011	Jiang et al, 2011
Usernames	**X**	**X**	**X**	**X**	
Links	**X**	**X**	**X**	**X**	
Mehrfach Iteration	**X**			**X**	**X**
Stopwords			**X**	**X**	
Lexikon				**X**	

Tabelle 1 Eingesetzte Filtermöglichkeiten bei ausgewählten Experimenten im Bereich Twitter Sentimentsmining

Kouloumpis et al haben bei der Filterung von Abkürzungen und Emoticons das Internet Lingo Dictionary verwendet sowie andere nicht näher spezifizierte Quellen (Kouloumpis et al., 2011). Bei der Suche nach diversen Slang- und Abkürzungswörterbücher fanden sich z. B. diese:

- http://www.netlingo.com (API auf Anfrage vorhanden)
- http://www.noslang.com
- http://www.internetslang.com
- http://www.urbandictionary.com (API auf Anfrage vorhanden)

sowie die Wikipedia mit folgenden hilfreichen Quellen:

- http://de.wikipedia.org/wiki/Liste_der_Abk%C3%BCrzungen
- http://de.wikipedia.org/wiki/Liste_der_Abk%C3%BCrzungen_(Netzjargon) oder
- http://de.wiktionary.org/wiki/Verzeichnis:ASCII-Smileys

Wirft man einen Blick auf diese vorhandenen Ressourcen, so zeigt sich, dass die Ressourcen über ein Webinterface zu nutzen sind, jedoch standardisierte Schnittstellen höchstens nur auf Nachfrage kostenpflichtig verfügbar sind. Auf der Website Internetslang.com oder Noslang.com können englischsprachige Abkürzungen und Wörter aus dem täglichen Sprachgebrauch abgerufen werden. So findet sich dort, dass die Abkürzung JLMK für „Just let me know“ [7] steht oder IJDK für „I just don’t know“[8]

Tabelle der Akronyme und Abkürzungen [Bearbeiten]

0–9 [Bearbeiten]

Abkürzung	Bedeutung
143	„I love you“
2F4U	„Too Fast For You“
2L8	„Too Late“
4U	„For you“
4YEO/FYEO	„For Your Eyes Only“

A [Bearbeiten]

Abkürzung	Bedeutung
AAMOF	„As A Matter Of Fact“
Acc	„Account“
ACK	„Acknowledgment“
AFAIC	„As Far As I'm Concerned“
AFAIK	„As Far As I Know“
AFAIR	„As Far As I Remember“
AFK	„Away from Keyboard“
AGF	„Assume good faith“
ASAP	„As Soon As Possible“
ASL (auch A/S/L)	„Age Sex Location“
ATM	„At The Moment“

Abbildung 7 Abkürzungen aus dem Netzjargon auf Wikipedia[9]

[77] http://www.internetslang.com/JLMK-meaning-definition.asp (abgerufen am 14.2.2013)

[8] http://www.internetslang.com/IJDK-meaning-definition.asp (abgerufen am 14.2.2013)

[9] Screenshot von http://de.wikipedia.org/wiki/Liste_von_Abk%C3%BCrzungen_(Netzjargon) (abgerufen am 14.2.2013)

Der Blick auf diese Lexika und Wörterbücher zeigen rasch, dass der automatisierte Zugang über Schnittstellen nicht einfach möglich ist und somit der kostenpflichtige Zugriff über API notwendig ist, oder ein eigenes Lexikon aufgebaut wird und dieses z. B. mit der Unterstützung von Scraping-Technologien gefüllt wird[10]. Die vorhandenen Listen auf der Wikipedia sind zumindest unter bestimmten Bedingungen frei verwendbar. Wie ein eigener Wortschatz aufgebaut wird, bzw. welche Anforderungen es an diesen gibt, dazu wurde vom Autor in der Literaturrecherche zum Zeitpunkt der Ausarbeitung der Seminararbeit und bei der Überarbeitung der selbigen leider keine Anhaltspunkte gefunden.

Stemming: Jiang et al führen an, dass sie in ihrer Arbeit einen Stemmer verwenden, welcher auf Basis einer Mapping-Tabelle funktioniert und über 20.000 Einträge verfügt (Jiang et al., 2011). Konkrete Namen führen die Autoren aber nicht an.

7.2 Werkzeuge und Quellen für die Merkmalsklassifizierung

Die folgenden Autoren haben hier laut den jeweiligen Artikeln folgende Werkzeuge zur Feststellung und Klassifizierung von Merkmalen bei Wörtern ausgewählt und eingesetzt (Tabelle 2):

[10] Die rechtliche Komponente wird in dieser Arbeit nicht beachtet.

	Go et al, 2009	Ritter et al, 2011	Pak et al, 2010	Maynard et al, 2012	Koulo-umpis et al, 2011	Jiang et al, 2011
Uni-Gram-me	**X**		**X**		**X**	
Bi-Gram-me	**X**		**X**		**X**	
Uni- und Bi-Gram-me	**X**					
Uni-Gramm mit POS-Tagging	**X**				**X**	
POS-Tagging		**X**		**X**		**X**

Tabelle 2 Eingesetzte Verfahren zur Merkmalsklassifizierung bei ausgewählten Experimenten im Bereich Twitter Sentimentsmining

Bei der Erstellung der N-Gramme verweist Pak et al darauf, dass diese bei einer Verneinung wie „no“ oder „not“, diese Verneinung mit dem Wort davor bzw. danach kombiniert haben, um die Genauigkeit der Klassifizierung zu verbessern (Pak & Paroubek, 2010).

Ritter et al haben in einem Artikel verschiedene POS-Tagger getestet und dabei einen eigenen POS-Tagger ins Rennen gesendet, welcher verglichen zu den bestehenden POS-Taggern für Tweets in englischer Sprache ein besseres Ergebnis erzielen konnte (Ritter et al., 2011). Dieser POS-Tagger ist frei auf Github verfügbar und wurde auf Tweets mit Datum- bzw. Zeitangaben für eine Kalenderapp trainiert.[11] Einen eigenen POS-Tagger für Twitter haben auch Gimpel et al entwickelt und frei zur Verfügung gestellt. Bei deren Experiment haben sie aber schlechtere Ergebnisse mit ihren Daten erhalten im Vergleich zum Stanford Tagger (Gimpel, Schneider, O’Connor, & Das, 2010).

Maynard et al haben auf eine adaptiere Version von ANNIE[12] zurückgegriffen und erste Tests durchgeführt. Das Feststellen von Named Entitities beschreiben die Autoren mit *„… is challenging and we are currently performing some separate experiments about this.“* (Maynard et al., 2012). Unter anderem auf Grund dieser Aussage schließt der Autor der Arbeit daraus, dass die dort beschriebenen Ergebnisse über

[11] https://github.com/aritter/twitter_nlp [4.12.2012]
[12] http://services.gate.ac.uk/annie/ [4.12.2012]

wenig Aussagekraft verfügen und es sich beim beschriebenen Projekt noch um einen prototypischen Status handelt.

7.3 Zusammenfassung

Auffällig ist beim Studieren der Artikel zum aktuellen Forschungsstand, dass viele Autoren die Datenaufbereitung eher im geringen Ausmaß vornehmen und sie sich hier auf einfachere Filter verlassen. Nur wenige Autoren nehmen Lexika oder Wörterbücher zu Hilfe, um die vorhandenen Texte anzureichern.

Dieses Kapitel zeigte teilweise vergleichend auf, welche Werkzeuge zur Verfügung stehen und auch eingesetzt werden. Bezeichnenderweise fand sich kein Framework mit einzelnen Komponenten, welches zur freien Verfügung steht bzw. testweise eingesetzt werden kann. Meist werden einzelne Komponenten frei angeboten und die Zusammenstellung zu einem funktionierendem Ganzen obliegt einem selbst. Auf welche Komponenten zurückgegriffen werden könnte zeigte dieses Kapitel in Auszügen.

8 Fazit

Die Herausforderungen die bei der verwendeten Sprache in den Microblogs entstehen ist in der Forschung bekannt und wird auch bereits durch Forschungsgruppen behandelt. Die verwendeten Ansätze ähneln sich hierbei, denn man trifft in den einzelnen wissenschaftlichen Herangehensweisen immer wieder auf ähnliche Muster. Das Wichtigste ist die Aufbereitung der Daten, die speziell bei der jugendlichen Sprache mit verkürzter Syntax bei Statusmeldungen noch ein wichtigerer Teil ist. Interessanterweise verweisen einige Autoren darauf, dass zwar sehr viele Rechtschreib- und Grammatikfehler entstehen bei den Statusmeldungen, doch der Autor entdeckte bei niemanden eine Komponente in deren Lösungen, die diese Fehler ausbessern würden. Komponenten die Abkürzungen voll ausschreiben, Mehrfach Iterationen auflöst und ähnliche Komponenten sind aber State of the Art und werden in der Regel von den meisten Wissenschaftlern in dieser Domäne angewendet.

Bei der Weiterverarbeitung der Daten in Richtung Feststellung des Sentiments finden sich interessanterweise unterschiedliche Ansichten und Herangehensweise. So setzen viele auf N-Gramme (in den verschiedensten Ausprägungen) und wenige auf POS-Tagger. Manche kombinieren auch beide Herangehensweisen. Was sich zeigte, dass herkömmliche POS-Tagger bei so kurzen Texten kaum Chancen besitzen auf ein akzeptables Ergebnis und so auch spezielle POS-Tagger für Twitter entwickelt und eingesetzt wurden.

Die bisherigen Ergebnisse zeigen, dass es bereits erste Ansätze gibt, aber es noch einiges an Entwicklungsarbeit in den verschiedensten Bereichen benötigt, um ein akzeptables Endergebnis zu erreichen.

Die Herausforderungen die sich aus der Literatur ableiten lassen gelten auch für andere Sprachen, wenn sich auch die Mehrheit der Literatur auf die englische Sprache bezieht und auf andere Sprachen in der Regel nicht eingeht. Wenn Werkzeuge wie ein

POS-Tagger zum Einsatz kommt, so sind hier spezielle POS-Tagger zu verwenden, wie es bspw. in dieser Arbeit mit dem STTS Tagset gezeigt wurde. Wird bspw. statt POS Tagging N-Gramme verwendet, so ist es notwendig, dass ein dementsprechender Korpus in der jeweiligen Sprache vorliegt.

Literaturverzeichnis

Bethard, S., Yu, H., & Thornton, A. (2006). *Extracting opinion propositions and opinion holders using syntactic and lexical cues.* Computing Attitude and Affect in Text: Theory and Applications, 125–141. Abgerufen von http://www.springerlink.com/index/G348K869J35R42P5.pdf

Connor, B. O., Smith, N. A., & Xing, E. P. (2008). *A Latent Variable Model for Geographic Lexical Variation.* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Facebook Inc. (2012). *One Billion Fact Sheet.* http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract (abgerufen am 16. Oktober 2012)

Frawley, W. J., Piatetsky-shapiro, G., & Matheus, C. J. (1992). *Knowledge Discovery in Databases: An Overview.* AI Magazine 13(3), 13(3), 57–70. Abgerufen von http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011

Gimpel, K., Schneider, N., O'Connor, B., & Das, D. (2010). *Part-of-speech tagging for twitter: Annotation, features, and experiments.* Abgerufen von http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA547371

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.* Abgerufen von http://cs.wmich.edu/~tllake/fileshare/TwitterDistantSupervision09.pdf

Groot, R. de. (2012). *Data Mining for Tweet Sentiment Classification.* Masterarbeit an der Universität Ütrecht. Abgerufen von http://igitur-archive.library.uu.nl/student-theses/2012-0824-200523/UUindex.html

Heyer Gerhard. (2011). *Text Mining: Wissensrohstoff Text.* https://docs.google.com/viewer?a=v&q=cache:gEJ04fQxQYQJ:wortschatz.uni-leipzig.de/~sbordag/TM07/TM01_TextWissenTM.pdf+&hl=de&gl=at&pid=bl&srcid=ADGEESiitkMirY8fqvMpYK_MLAZc5gTcgEY1kgrnLyZDST-k-YD-D9H3ePdeqWYkftuvKh7Y4pCfO27vgYmiEZKxaJ96NunokXKsB002aR0My6Sh1tI8GPHYXk7OF1hCuVyf7Kt-kCwt&sig=AHIEtbRpoYWwrZS5jKFJMPDefXclIMWjKA (abgerufen am 18. November 2012)

Institut für deutsche Sprache. (2012a). *Morphosyntaktische Annotationen.* http://www.ids-mannheim.de/cosmas2/projekt/referenz/annotationen.html (abgerufen am 18. November 2012)

Institut für deutsche Sprache. (2012b). *Stuttgart-Tübingen-Tagset (STTS).* http://www.ids-mannheim.de/cosmas2/projekt/referenz/stts/ (abgerufen am 18. November 2012)

Institut für Technikfolgen-Abschätzung der österreichischen Akademie der Wissenschaften. (2009). *Microblogging und die Wissenschaft. Das Beispiel Twitter.* http://epub.oeaw.ac.at/ita/ita-projektberichte/d2-2a52-4.pdf (abgerufen am 18. November 2012)

Jiang, L., Yu, M., & Zhou, M. (2011). *Target-dependent twitter sentiment classification.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 151–160. Abgerufen von http://newdesign.aclweb.org/anthology-new/P/P11/P11-1016.pdf

Kalisch, F. (2012). *Opinion Mining.* Seminararbeit an der Hochschule Furtwangen. Abgerufen von www.florian-kalisch.de/projects/om/paper-opinion-mining.pdf (abgerufen am 11. November 2012)

Kouloumpis, E., Wilson, T., & Moore, J. (2011). *Twitter sentiment analysis: The good the bad and the omg.* Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538–541. Abgerufen von http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2857/3251

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). *Recognizing Named Entities in Tweets.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), (2008), 359–367.

Lui, M., & Baldwin, T. (2011). *Cross-domain Feature Selection for Language Identification.* Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), (1967), 553–561. Abgerufen von http://newdesign.aclweb.org/anthology/I/I11/I11-1062.pdf

Manning, C. (2011). *Part-of-speech tagging from 97% to 100%: is it time for some linguistics?* Computational Linguistics and Intelligent Text Processing, 171–189. Abgerufen von http://www.springerlink.com/index/Y2H44828N1826661.pdf

Maynard, D., Bontcheva, K., & Rout, D. (2012). *Challenges in developing opinion mining tools for social media.* Proceedings of@ NLP can u tag# usergeneratedcontent.
Abgerufen von http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf

McCombs, M. M. E., & Shaw, D. D. L. (1972). *The Agenda-Setting Function of Mass Media. Public opinion quarterly, 36(2)*, 176–187.
Abgerufen von http://poq.oxfordjournals.org/content/36/2/176.short

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., et al. (2011). *Quantitative analysis of culture using millions of digitized books.* Science (New York, N.Y.), 331(6014), 176–82.
doi:10.1126/science.1199644

Moraldo, S. M. (2006). *das Leben in 140 Zeichen... heisst Twitter:-*). sprachverein.ch.
http://www.sprachverein.ch/moraldo.pdf (abgerufen am 23. November 2012)

Owoputi, O., O'Connor, B., & Dyer, C. (2012). *Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances*.
http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.tr12.pdf (abgerufen am 17. Oktober 2012)

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* Proceedings of LREC, 1320–1326.
Abgerufen von http://deepthoughtinc.com/wp-content/uploads/2011/01/Twitter-as-a-Corpus-for-Sentiment-Analysis-and-Opinion-Mining.pdf

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), (July), 79–86.
Abgerufen von http://dl.acm.org/citation.cfm?id=1118704

Rainie, L., Brenner, J., & Purcell, K. (2012). *Photos and Videos as Social Currency Online shares of men and women.*
http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_OnlineLifeinPictures.pdf (abgerufen am 17. Oktober 2012)

Ritter, A., Clark, S., & Etzioni, O. (2011). *Named entity recognition in tweets: an experimental study.* Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1524–1534.
Abgerufen von http://dl.acm.org/citation.cfm?id=2145595

Ritter, A., Etzioni, O., & Clark, S. (2012). *Open domain event extraction from twitter.* Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12, 1104.

Statista.de. (2012). *Twitter - Anzahl der monatlich aktiven Nutzer 2012 | Statistik.* http://de.statista.com/statistik/daten/studie/232401/umfrage/monatlich-aktive-nutzer-von-twitter-weltweit-zeitreihe/ (abgerufen am 17. Oktober 2012)

Statista.de, & comScore. (2010). *Altersverteilung der Nutzer von Twitter in Europa im Dezember 2010.* http://de.statista.com/statistik/daten/studie/193015/umfrage/nutzer-von-twitter-in-europa-nach-alter/ (abgerufen am 17. Oktober 2012)

TeleRing. (2011). *So telefoniert und smst Österreichs Jugend.* http://www.ots.at/presseaussendung/OTS_20110831_OTS0166/studie-so-telefoniert-und-smst-oesterreichs-jugend-nicht-nur-die-liebe-geht-ueber-das-handy-bild (abgerufen am 18. Oktober 2012)

Twitter Inc. (2011). *Twitter Blog: One hundred million voices.* http://blog.twitter.com/2011/09/one-hundred-million-voices.html (abgerufen am 16. Oktober 2012)

Kommentierte Literaturliste

Bethard, S., Yu, H., & Thornton, A. (2006). Extracting opinion propositions and opinion holders using syntactic and lexical cues.

Opinion Mining Extraction. Erklärt Vorgehensweise und Ergebnisse bei einem Korpus mit > 5000 Sätzen. Sieht keine Notwendigkeit des zusätzliches Einsatzes eines POS-Taggers, da die Qualität nicht verbessert wird.

Gimpel, K., Schneider, N., O'Connor, B., & Das, D. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments

Entwickelten einen eigenen POS-Tagger und stellten diesen zur freien Verfügung. Resultat waren aber schlechtere Ergebnisse. Erste Versuche mit einem eigenen POS-Tagger aber kein Quantensprung. Weitere Entwicklung?

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision.

Paper zum Thema Opinion Mining, welche s anführt, dass Sätze mit Emoticons sehr geeignet für Trainingsdaten für die div. Algorithmen sind. Besitzen Erfahrung in der Datenaufbereitung für Opinion Mining

Groot, R. de. (2012). Data Mining for Tweet Sentiment Classification.

Masterarbeit über die Vorgehensweise zum Opionion Mining. Interessant, tw. allgemeiner gehalten. Geht nicht so in die Tiefe wie die div. Papers

Heyer Gerhard. (2011). Text Mining: Wissensrohstoff Text.

LVA-Folien zum Thema. Wohl einer der Experten im deutschsprachigem Raum. Auch Buch vorhanden (ISBN 978-3937137308)

Jiang, L., Yu, M., & Zhou, M. (2011). Target-dependent twitter sentiment classification.

Das Paper beschreibt den Ansatz von Jiang et al zum Opinion Mining, welcher abweicht von den anderen Autoren. Keine näheren Angaben zum konkreten Einsatz, daher nur bedingt zu gebrauchen.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg.

Weiteres Paper zum Opinion Mining allgemein. Beschreibung der Vorgehensweise und auch der Einsatz von Lexika. Liest sich kompetent.

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets.

Paper befasst sich ausschließlich zum Thema NER und deren Ansätze. Autoren empfehlen einen kombinierten Ansatz auf KNN und CRF

Lui, M., & Baldwin, T. (2011). Cross-domain Feature Selection for Language Identification.

Es wird ein eigener Ansatz zur Identifikation der Sprache auf Basis eines vorgegebenen Textes vorgestellt. Ebenfalls werden andere Möglichkeiten vorgestellt. Interessant vor allem, weil auch andere Möglichkeiten vorgestellt werden, wie eine Sprache erkannt werden kann.

Manning, C. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?

Blick hinter die Geschichte und bisherigen Ergebnisse in Hinblick auf Präzision von gängigen POS-Taggern und Vorstellung des eigenen Konstrukts. Interessant zu

lesen, da aufkommt, dass die POS-Tagger doch nicht so genau arbeiten, speziell auf Satzebene. Der vorgestellte neue POS-Tagger ist angeblich um eine Spur besser.

Moraldo, S. M. (2006). das Leben in 140 Zeichen... heisst Twitter:-)

Der Sprachwissenschafter stellt die Besonderheiten der Texte auf Twitter vor und analysiert diese in verschiedene Richtungen. Sehr gute Zusammenfassung über die Herausforderungen aus der Sicht der Sprache.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining.

Opinion Mining Paper welches die Vorgehensweise bis zur Analyse des Sentiments beschreibt. Methode und Ergebnisse werden mE gut dargestellt.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques.

Ein weiteres Paper über die Vorgehensweise bei der Sentimentsanalyse von Tweets. Beschreiben nur ihre ersten Versuche in diesem Umfeld.

Ritter, A., Clark, S., & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study.

Autoren zeigen die Vorgehensweise der Analyse und warum sie so vorgehen. Autoren sind Experten auf diesem Gebiet und zeigen dies auch.

Ritter, A., Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter.

Autoren zeigen in diesem weiteren Paper ein Experiment, welches sie auch öffentlich zur Verfügung stellen zur Weiterentwicklung. Interessant, aber fokussiert auf lediglich eine Domäne. Übertragbarkeit der Erkenntnisse auf andere Domänen mehr als fraglich.

Anhang

Researchmap

Keywords	Personen	Universitäten	Systeme
Twitter NLP Opinion Mining POS Tagging	Olutobi Owoputi, Brendan O'Connor, Kevin Gimpel, Nathan Schneider, Chris Dyer, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah Smith.	ARK Social Media Research, Carnegie Mellon University.	ark-tweet-nlp http://code.google.com/p/ark-tweet-nlp/downloads/list
NER Opinion Mining POS-Tagging N-Gram	Ritter, Alan Clark, Sam Etzioni, Oren	University of Washington	Twitter_nlp https://github.com/aritter/twitter_nlp http://statuscalendar.cs.washington.edu/
Data Mining Tweet Sentiment Classification	Groot, R de	Universität Ütrecht	MasterThesis
NLP Allgemein		Google	http://research.google.com/pubs/NaturalLanguageProcessing.html
NER	Xiaohua Liu Shaodian	Shanghai Jiao Tong University	

	Zhang-Furu Wei Ming Zhou		

Tagcloud

Abbildung 8 Eigene Abbildung - erstellt mit Wordle.net