Automatische Verarbeitung deutschsprachiger Tweets: Eine Fallstudie

1 Einführung

TODO: DIE BESONDERHEITEN DES TWEET PROCESSING

2 Zielsetzung

Das Projekt "Diskurse in Social Media" untersucht aus kommunikationswissenschaftlicher Perspektive den Verlauf von politischen Diskursen in drei verschiedenen Social Media: Twitter, Facebook und Blogs. Es wird untersucht, inwieweit sich zwischen diesen drei Medien Unterschiede finden lassen im Hinblick auf

- den zeitlichen Verlauf von Debatten: Wie entwickelt sich das Volumen einer Diskussion?
- die Dialogizität der Diskurse: Wie gehen Teilnehmer auf die Beiträge anderer Teilnehmer ein?
- die Meinungsführerschaft: Sind bestimmte Akteure in den Diskussionen "tonangebend"?
- die Themen: Welche Aspekte des Themenkreises werden diskutiert?
- die Meinung: Welche Stimmungslage kommt in einer Diskussion zum Ausdruck?

Die qualitativ hochwertige Analyse dieser Fragestellungen setzt das sachkundige menschliche Urteil voraus, das heißt: am Ende der Bemühungen steht eine Inhaltsanalyse durch ausgebildete Analysten. Die Projektpartner aus der Wirtschaftsinformatik und der Computerlinguistik sollen diesen Prozess

aber maßgeblich unterstützen, um bei einem relativ umfangreichen Datensatz die menschliche Analyse auf die relevanten Beiträge konzentrieren zu können. Als erstes Arbeitskorpus wurden dazu Daten ausgewählt, die im Jahr 2011 über einen Zeitraum von zwei Wochen hinweg (TODO: stimmt das? - wenn die Angabe 12.12-17.02 im Dateinamen des Korpus den Zeitraum des Textsammelns bedeutet, so dürften es gut 2 Monate sein) das Stichwort 'Wulff' beinhalten, also höchstwahrscheinlich Äußerungen zur Affäre um den seinerzeitigen Bundespräsidenten beinhalten. Im vorliegenden Beitrag beschränken wir uns auf die Twitter-Daten, das sind 119 455 einzelne Tweets, von denen wir zunächst 28 818 als Duplikate idetifiziert haben; somit verbleibt eine Grundmenge von 90 637 zu verarbeitenden Tweets.

Im Projekt sind die Wirtschaftsinformatiker für die Beschaffung der Datensätze und die Konstruktion von Graphstrukturen zuständig, die die Verweise zwischen Diskussionsbeiträgen abbilden. Der Computerlinguistik obliegt die inhaltliche Analyse der einzelnen Beiträge im Hinblick auf behandelte Themen und die Stimmungslage: Es wird eine "Vorsortierung" der Beiträge durchgeführt, um im Idealfall die Aufgabe der Analysten auf eine zügige Durchsicht beschränken zu können. Die Vorsortierung erfolgt anhand drier Dimensionen:

- Themenklassifikation: Durch unüberwachte Verfahren wird ein Clustering von Beiträgen im Hinblick auf (Unter-) Themen vorgenommen.
  Hier einige Beispiele aus dem Wulff-Korpus, die das Adressieren verschiedener Themen illustrieren: TODO

- Sentimentklassifikation: Für jeden Einzelbeitrag soll erkannt werden, ob eine positive, negative oder neutrale Haltung ausgedrückt wird.
  Korpus-Beispiele: TODO
  Zusätzlich soll im Falle von Verweisen auf andere Beiträge festgestellt werden, ob diese zustimmend, kritisch, oder neutral ausfallen. Korpusbeispiele: TODO

- Diskursqualitätklassifikation: Textbeiträge können inhaltlich fundiert sein und das Potenzial haben, eine Diskussion fruchtbar voranzubringen, oder lediglich kurze "Einwürfe" darstellen. Korpusbeispiele: TODO

3 Die Pipeline zur Vorverarbeitung

TODO:

3.1 Sprachidentifikation

Die von den Wirtschaftsinformatikern zusammengestellten Daten enthalten noch tweets, die nicht in deutscher Sprache abgefasst sind. Um diese zu filtern, haben wir mit drei *off-the-shelf* Werkzeugen zur Sprachidentifikation experimentiert.

- Google Language Identifier. TODO: weiß man wie er arbeitet? : Sprachmodelle mit pro Sprache 300 Trigrammen; 68 Sprachen
  Aber: Wie erkennt er Griechisch, Koreanisch, Japanisch und Chinesisch, wenn er keine Trigramme dafür hat???
- lang-ident
- textcat

TODO: Vergleichende Evaluation: Methode, Durchführung.
wüßte mal gern ob's klappt

3.2 Satzgrenzenerkennung

Die generelle Arbeitsweise eines "sentence splitters" besteht darin, Punkte am Ende eines Wortes dahingehend zu disambiguieren, ob es sich um einen Satzende-Punkt oder den Bestandteil einer Abkürzung (wie z.B. in "Nr. 3"), einer Datumsangabe o.ä. handelt.

TODO: Besonderheiten bei Tweets?

3.3 Normalisierung und Tokenisierung

3.3.1 Zu behandelnde Phänomene

TODO: Klassifizierte Liste von Dingen, die man behandeln muss

3.3.2 Vorgehen

3.4 Part-of-speech tagging

TODO: Tree Tagger versus TNT.

4 Inhaltsklassifikation / Auswertung

4.1 Koreferenz

Für eine genauere Analyse des Tweet-Inhalts auf der Satzebene soll eine Koreferenzresolution vorgenommen werden.

TODO:

- Wieviele Pronomen finden wir?
- Wieviele sonstige Korefs?
- Wie ist die Performanz von PoCoRes?
- Performanz auf normalisiertem Input?
- Was sind die typischen Fehler?
- Perspektive: wie weiter? Ist twitter-Koref einfacher als generelle Koref? Was dürfte ein vielversprechendes Verfahren sein?

4.2 "Off-topic"

Die Datenerhebung geschah im Wesentlichen durch keyword matching: Beinhaltet ein tweet das Wort 'Wulff'? Dadurch können gelegentlich tweets in die Suchmenge gelangen, die thematisch nicht relevant sind, weil sie sich auf eine andere Person gleichen Namens beziehen.
TODO: Beispiel
TODO: Kommt das häufig vor?
TODO: Wie gehen wir damit um?

4.3 Subtopiks

4.4 Sentiment

4.5 Diskursqualität

5 Zusammenfassung