TITEL: Sentiment-Analyse – Algorithmus oder Stichprobe?

Autoren: Guido Moeser[1,2], Heiko Moryson[1]

[1] Wiesbaden R Users Group
[2] masem research institute, Wiesbaden

[Ziel: kleiner Vortrag, 15 bis 30 min.]

ABSTRACT:

Einleitung

Hintergrund:
Mit Sentiment-Analyse wird die Stimmungsanalyse von unstrukturierten Texten bezeichnet. Dabei werden aufgrund der hohen Anzahl an Texten zunehmend Algorithmen eingesetzt, um deren Sentiment zu bestimmen. Die Genauigkeit (engl. Accuracy) dieser automatischen Verfahren liegt jedoch meist zwischen 60 und 70 Prozent, was aber im Umkehrschluss bedeutet, dass die Verfahren 30 bis 40 Prozent der Texte nicht korrekt klassifizieren.
Neben der automatischen Klassifizierung aller Texte stellt die Ziehung einer Stichprobe aus der Gesamtheit und die manuelle Kodierung des Sentiments der Texte in der Stichprobe eine verbreitete Methode dar, um eine höhere Genauigkeit zu erreichen. Jedoch tritt hier ein Stichprobenfehler auf, d.h. ist die Stichprobe zu klein gewählt, so stimmt das Ergebnis in der Stichprobe nicht mit dem in der Grundgesamtheit überein.

Ziel:
Die beiden Verfahren sollen vorgestellt werden. Es soll überprüft werden, wann sich welcher Ansatz eignet. Dabei soll ein möglicher Trade-Off zwischen den Kosten der manuellen Klassifizierung auf Stichprobenbasis und der Genauigkeit bei Nutzung eines Algorithmus untersucht und dargestellt werden. Es soll gezeigt werden, inwieweit sich das Softwarepaket R für diese Analysen eignet.

Methode

Daten:
Zur Analyse werden englischsprachige Tweets des Dienstes Twitter genutzt. Über 5000 Tweets zu einem Thema werden hinsichtlich des Sentiments manuell kodiert. Dabei werden die Tweets eingeteilt in: (1) positive Stimmung, (2) negative Stimmung, (3) neutrale Stimmung, (4) keine Stimmung und (5) irrelevant für die Fragestellung.

Methode:
Die Statistik-Software *R* wird für die durchzuführenden Analysen eingesetzt. Zur automatischen Klassifizierung wird unter anderem das Text-Mining-Paket *tm* herangezogen. Geeignete Stimmungswörterbücher werden genutzt. Verschiedene Algorithmen werden ausprobiert, um eine Klassifizierung vorzunehmen. Die Genauigkeit wird festgestellt durch den Vergleich der manuellen Kodierung mit der Kodierung durch die Algorithmen.

Die Analyse mittels Stichprobe erfolgt mittels Resampling-Ansatz: Dabei werden Stichproben mit ansteigender Größe aus dem Gesamtdatensatz gezogen. Die Sentimentanteile der Stichproben werden mit den Sentimentanteilen des Gesamtdatensatzes verglichen.

Um eine höhere Praxisnähe zu erreichen, sollen – wenn möglich - die Analysen Datumsbezogen ausgeführt werden, d.h. die zu einem Tag zugehörigen Tweets werden entsprechend den beiden Methoden ausgewertet. So lassen sich auch Stimmungsverläufe über die Zeit ermitteln.

[335 Wörter]

Wiesbaden, 16. April 2013