CAS Information Engineering 2019.12/2020.01

Leistungsnachweis Information Retrieval

Sentiment Analyse von New York Times Artikeln und Kommentare

Abgabtermin: 12.01.2020

Autoren: Simon Stäheli, Simon Würsten

Inhaltsverzeichnis

1 Ausgangslage

Als Leistungsnachweis für das Modul «Information Retrieval» im CAS Information Engineering soll in einem Projekt die erlernten Methoden auf ein selbst ausgewähltes Datenset angewendet werden. Das Ergebnis zur Analyse umfasst eine mindestens 4 seitige Dokumentation und eine Präsentation am letzten Tag des CAS.

Wir haben uns für ein Datenset auf Kaggle entschieden, welches online publizierte Zeitungsartikel und deren Kommentare der New York Times vom Zeitraum Januar - Mai 2017 und Januar-April 2018 umfasst. Für unser Projekt betrachten wir den Zeitraum Januar-April 2018. Das Datenset ist in zwei CSV-Files unterteilt: eine Datei mit Informationen über die Artikeln und eine Datei mit Informationen über die Kommentare zu den Artikeln.

Wir verfolgen zwei Ziele:

- Sentiment der einzelnen Kommentare bestimmen
- Analyse: Rufen Artikel desselben Autors immer wieder ähnliche Sentiments und Keywords hervor?

Das Vorgehen ist wie folgt:

1. Artikelinhalt downloaden (Web Crawler)
2. Sätze tokenisieren
3. Satz- und Sonderzeichen sowie Zahlen entfernen
4. Wörter tokenisieren
5. Stoppwörter entfernen
6. Weitere Funktionen (Sentiments, Group By, Sum, …)

Nachfolgend zwei Screenshots mit den wichtigsten Spalten der Kommentare sowie dem Text (‘commentBody’ bzw. ‘words_clean’) vor und nach dem Tokenisieren und Entfernen der Stoppwörter.

	createDate_date	userID	commentBody
0	2018-04-24	46566740.0	How could the league possibly refuse this offe...
1	2018-04-24	64324866.0	So then the execs can be like "yeah...we will ...
2	2018-04-24	78105093.0	I would not want to play chess against these c...
3	2018-04-24	81939618.0	Could the cheerleaders join the Actors' Equity...
4	2018-04-24	58642997.0	Seeking conclusions which support preconceived...

Abbildung 1: Kommentar-Datenset vor der Bearbeitung

	createDate_date	userID	words_clean
0	2018-04-24	46566740.0	[could, league, possibly, refuse, offer]
1	2018-04-24	64324866.0	[execs, like, yeahwe, sit, listen, nothing, su...
2	2018-04-24	78105093.0	[would, want, play, chess, cheerleaders, lawye...
3	2018-04-24	81939618.0	[could, cheerleaders, join, actors, equity, as...
4	2018-04-24	58642997.0	[seeking, conclusions, support, preconceived, ...

Abbildung 2: Kommentar-Datenset nach der Bearbeitung

2 Herausforderungen

Die Daten zu den Artikeln enthielten nicht den eigentlichen Text des Artikels. Daher musste mit einem Web Crawler der Text geholt werden. Dank der vorhandenen Artikel-URL konnten die Artikel eindeutig identifiziert und heruntergeladen werden. Der individuelle Web Crawler wurde als Teil dieses Projektes implementiert. Zudem besteht zu diesem Datensatz keine Version mit bereits gelabelten Sentiments (negativ, neutral, positiv) pro Artikel oder Kommentar. Dies mussten wir selbstständig durchführen.

Bei den Kommentaren stellte die Datenmenge eine Herausforderung dar. Dieses Datenset umfasste mehr als 930'000 Kommentare bzw. Zeilen. Das Tokenisieren führte bei den Kommentaren zu ca. 64.6 Mio. Wörter (inkl. Stoppwörter) bzw. ca. 33.3 Mio. ohne Stoppwörter. Es waren demnach ca. 48% der Wörter Stoppwörter. Das Vorgehen dazu und die Funktionen für die weitere Verarbeitung dauerten lokal daher oftmals mehrere Minuten (> 10 Min oder gar nicht praktikabel.). So wurde bspw. nur vom häufigsten Wort die Anzahl über den gesamten Zeitraum dargestellt und Bigramme über den gesamten Korpus konnten nicht durchgeführt werden.

Als besonders langsam hat sich das Suchen von einem bestimmten Wort erwiesen. Hierzu muss einerseits das Data Frame Zeile für Zeile durchlaufen, andererseits pro Zeile die einzelnen Elemente der Liste mit den tokenisierten Wörtern gefiltert werden. Da Python für uns immer noch relativ neu ist, gibt es bestimmt effizientere Methoden als die von uns oft benutzten for-Schleifen. Der Umgang mit Listenspalten eines Dataframes, Groupby's oder Plots ist je nach notwendiger Komplexität allgemein zeitaufwendig.

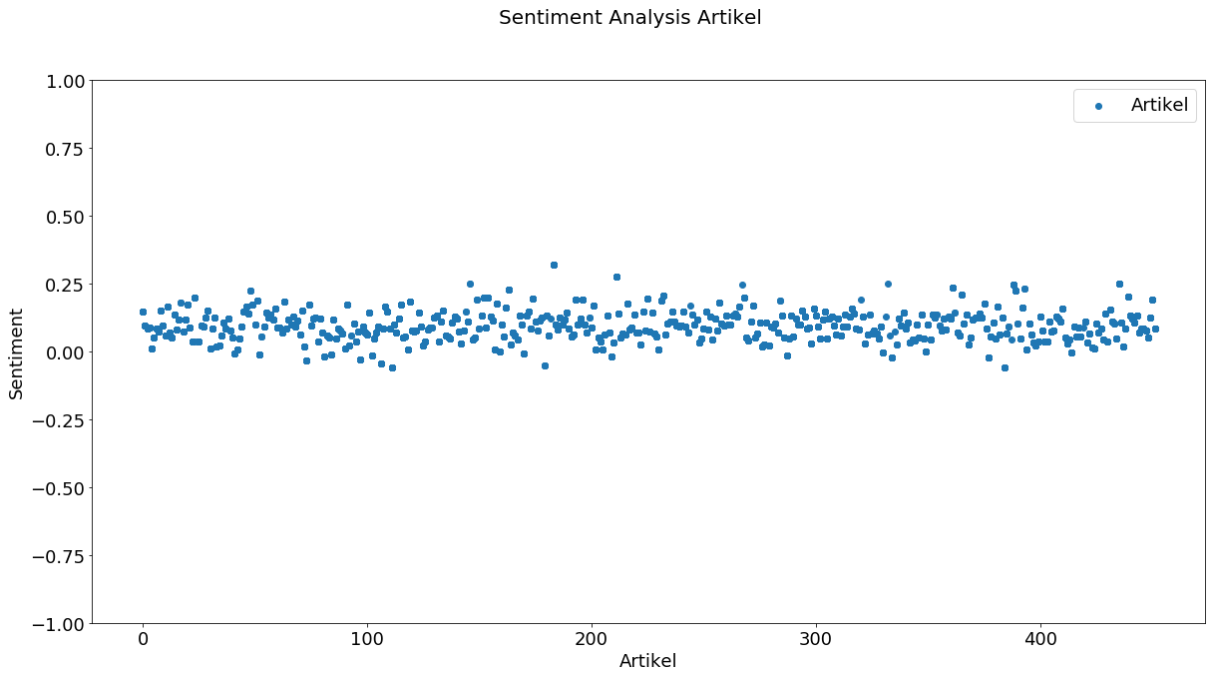
3 Analyse

3.1 Artikel

Die Autoren der Artikel sind nicht immer bekannt. Teilweise werden als Autoren Dinge ‘By THE EDITORIAL BOARD’ oder ‘Compiled by INSIDER STAFF’ wie angegeben. Dies erschwert die ursprüngliche Idee eine Sentiment-Analyse auf Basis der verschiedenen Artikel von Autoren vorzunehmen.

3.1.1 Sentiments

Die Analyse der Artikel Sentiments hat gezeigt, dass die Autoren der NYT relativ sachliche Texte schreiben. 95% der Artikel befinden sich im Band 0 – 0.25 wobei 1 ein sehr positives Sentiment darstellt und -1 ein sehr negatives.



3.2 Kommentare

Die Stoppwörter wurden für die nachfolgenden Analysen jeweils immer ausgeschlossen. Da für die Kommentare das Erstellungsdatum verwendet wurde, korreliert dieses nicht immer mit dem Publikationsdatum des Artikels. Daher sind auch noch Kommentare im Mai vorhanden, obwohl der Datensatz nur Januar-April abdeckt.

3.2.1 Verteilung Anzahl Wörter pro Kommentar

Es zeigt sich, dass die Verteilung sowohl pro Monat wie auch über alle Monate rechtsschief ist, wobei der Tail, bis auf den einen Ausreisser, nicht allzu lang ist. Ausserdem verbleiben 50% der Personen nach Abzug der Stoppwörter bei ca. 10-50 Wörter pro Kommentar. Interessanterweise gibt es eine kleine Erhebung bei ca. 125 Wörtern pro Kommentar.

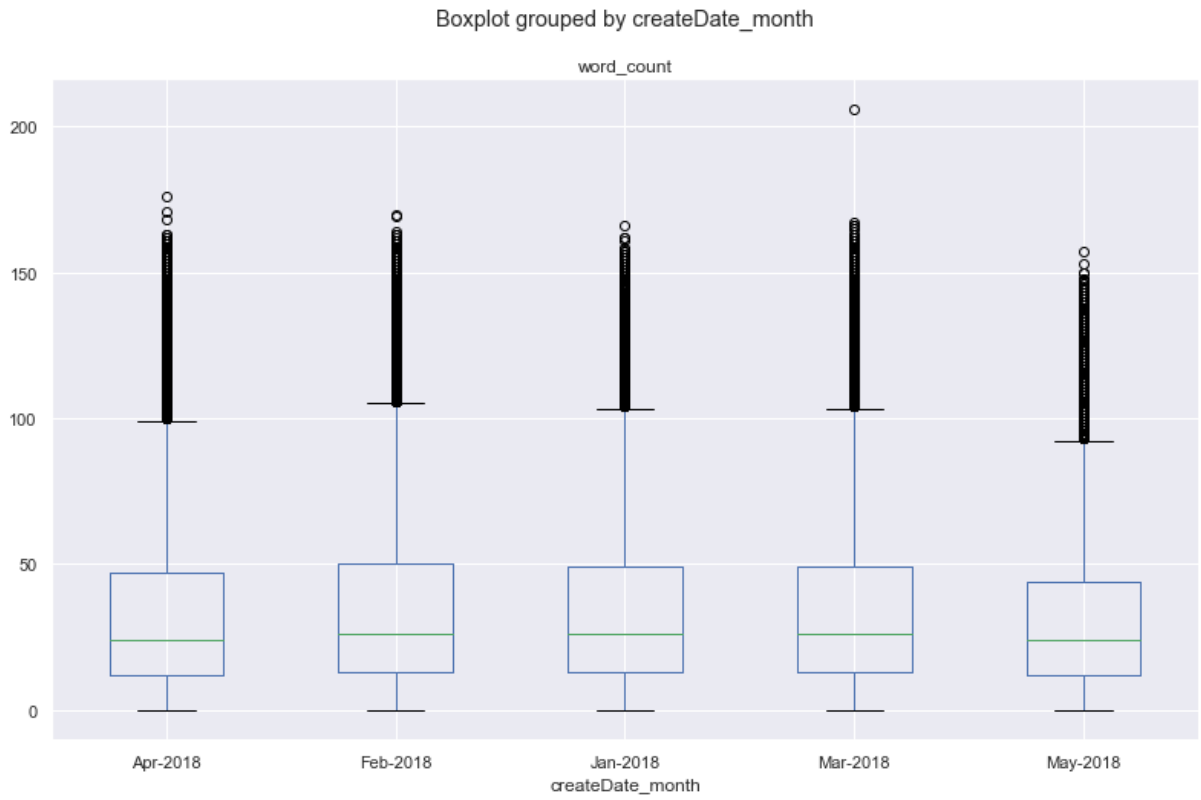


Abbildung 3: Verteilung Anzahl Wörter pro Kommentar (ohne Stoppwörter) pro Monat

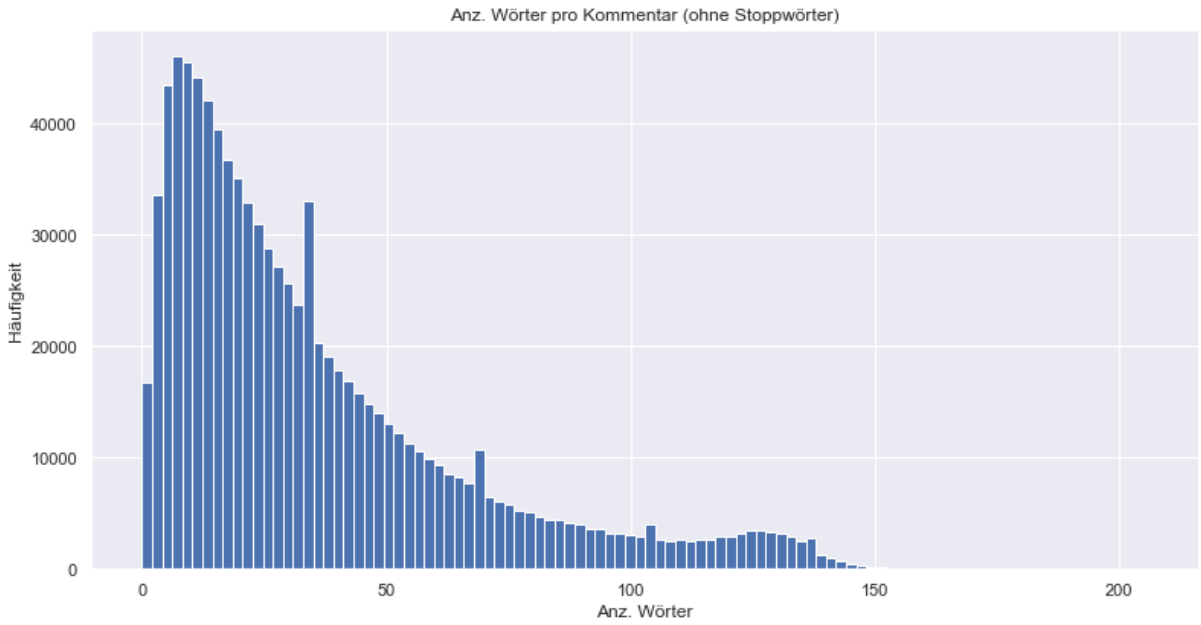


Abbildung 4: Verteilung Anzahl Wörter pro Kommentar (ohne Stoppwörter) alle Monate

3.2.2 Häufigste Wörter

Während der Zeitdauer von Januar – April 2018 scheint die User insbesondere ‘trump’ zu beschäftigen. Pro Monat betrachtet entspricht die Rangfolge der Wörter mehrheitlich dem Bild über den gesamten Zeitraum.

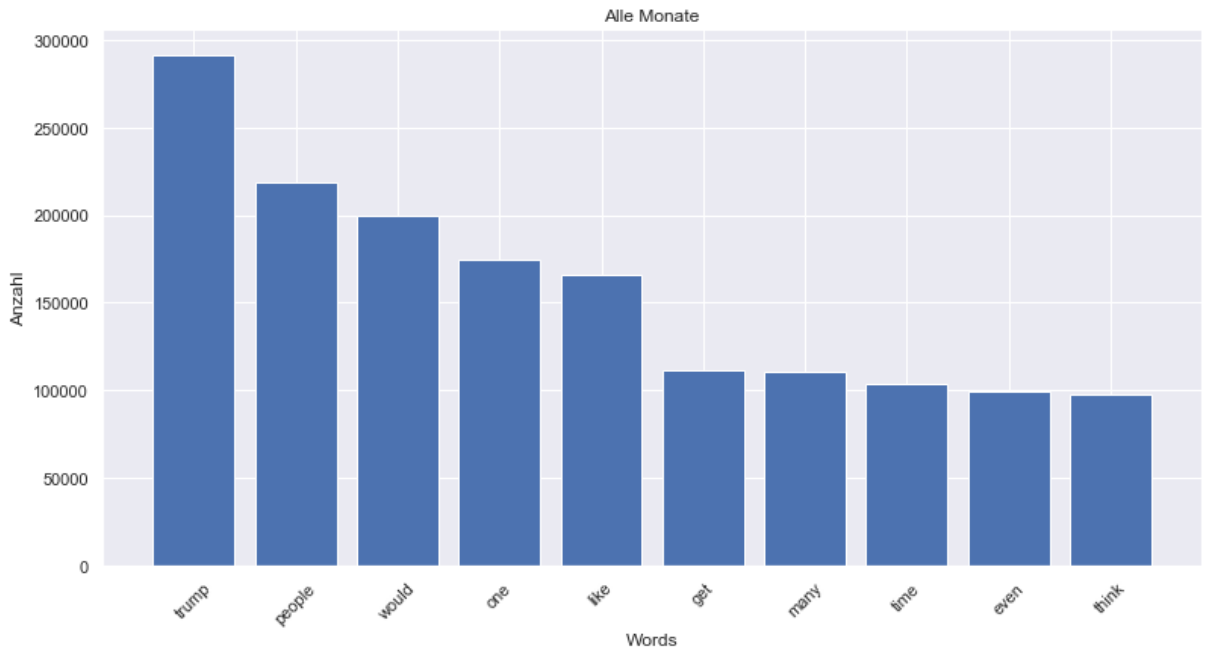


Abbildung 5: häufigste Wörter über gesamten Zeitraum

Farblich hervorgehobene Wörter kennzeichnen Abweichungen gegenüber den anderen Monaten im entsprechenden Rang. Es fällt auf, dass im Mai oft ein anderes Wort am meisten benutzt wurde; jedoch sind für den Mai am wenigsten Kommentare vorhanden, weshalb die Volatilität grösser ist. Bei Rang 6 und Rang 7 ist ‘get’ und ‘many’ für zwei aufeinanderfolgenden Monaten vertauscht.

Rang	Januar 2018	Februar 2018	März 2018	April 2018	Mai 2019
1	trump	trump	trump	trump	trump
2	people	people	people	people	would
3	would	would	would	would	questions
4	one	one	one	one	mueller
5	like	like	like	like	tax
6	get	get	many	many	people
7	many	many	get	get	one
8	time	time	time	time	like
9	president	even	even	even	get
10	think	gun	think	think	know

Tabelle 1: Top10 häufigste Wörter pro Monat

Abschliessend ist die Anzahl Vorkommen des häufigstes Wortes ‘trump’ über den gesamten Zeitraum dargestellt. Es scheint eine wöchentliche Saisonalität zu geben. Aufgrund der langsamen Performance wurde darauf verzichtet, andere häufige Wörter über die Zeit darzustellen.

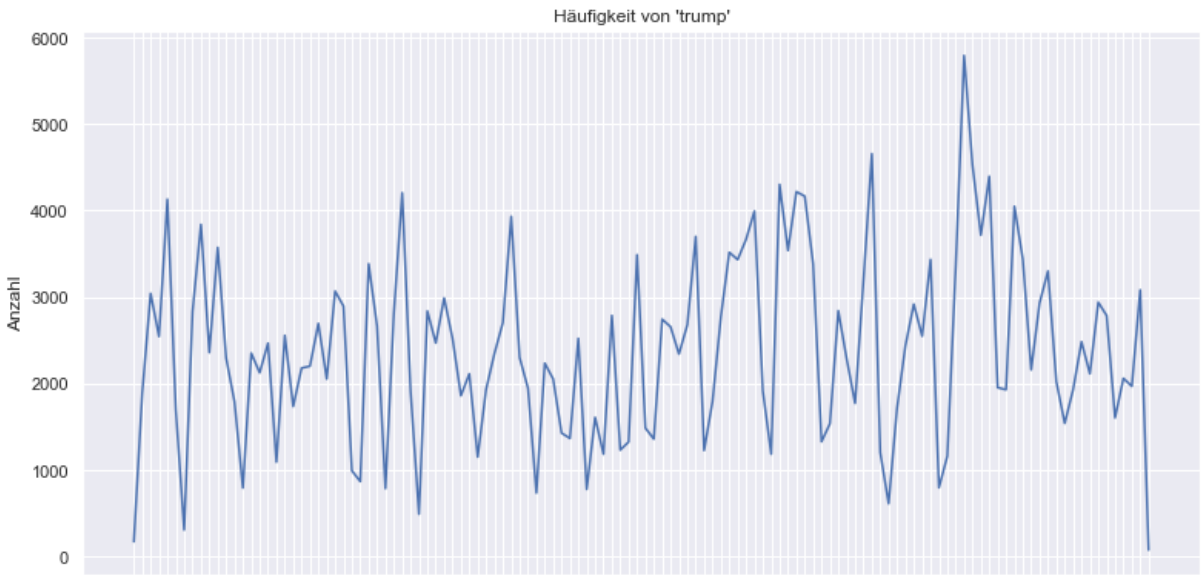


Abbildung 6: Ânzahl häufigstes Wort 'trump' über den gesamten Zeitraum

3.2.3 Bi-gramme

Im Gegensatz zu den häufigsten Wörtern unterscheiden sich die Bigramme innerhalb der Monate. Wie oben einleitend bereits erwähnt, konnten aufgrund von Performance-Schwierigkeiten keine Bigramme über alle Monate hinweg erstellt werden. Dennoch zeigt sich, dass die Kommentare insbesondere politische Themen behandeln und dazu in verschiedenen Wortzusammensetzungen auf die Regierung der USA Bezug nehmen (bspw. ‘white house’, ‘united states’, ‘mr trump’, ‘donald trump’). Damals aktuelle Themen in den USA, die auch bei uns in den Medien erschienen sind und diskutiert wurden, sind ebenfalls erkennbar, wie zum Beispiel die Gesundheitsvorsorge (‘health care’), der Einfluss von Social Media vermutlich auf die Wahlen (‘social media’) oder die Regulierung des Waffenbesitzes (‘gun control’)

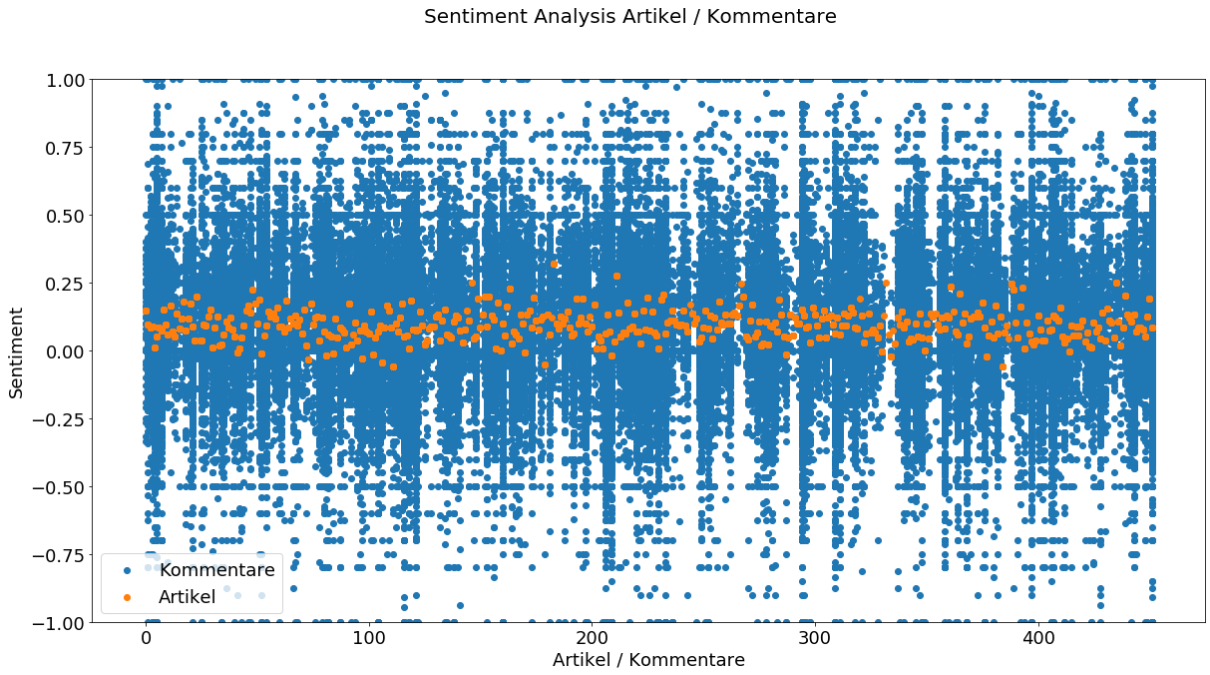
Rang	Januar 2018	Februar 2018	März 2018	April 2018	Mai 2018
1	white house	white house	white house	donald trump	tax cuts
2	donald trump	united states	united states	white house	tax cut
3	united states	donald trump	donald trump	united states	mr mueller
4	new york	gun control	years ago	new york	legal team
5	republican party	years ago	new york	years ago	white house
6	years ago	republican party	mr trump	mr trump	united states

7	health care	social media	north korea	republican party	special counsel
8	mr trump	second amendment	social media	many people	muellers team
9	president trump	many people	gun control	fox news	mr trump
10	many people	high school	many people	health care	new york

Tabelle 2: Top10 Bigramme nach Monat

3.2.4 Sentiments

Ein Vergleich der Sentiments der Artikel mit den Sentiment der Kommentare hat gezeigt, dass die NYT-Leser sehr ausgeglichen kommentieren.Statistisch gesehen gibt es leicht mehr positive Kommentare als negative, gesamthaft fällt dies aber nicht auf.



Desweitern wurden folgende Beziehungen hinsichtlich eines Sentiment-Trends weiter untersucht. Dabei konnten aber leider keine Trends festgestellt werden.

- Welche Auswirkung haben Artikel mit sehr positivem Sentiment (> 0.25) auf deren Kommentare
- Welche Auswirkungen haben Kommentierer grundsätlich bzw. die Sentiments deren Kommentare im Vergleich zum Sentimen des Artikels
- Welche Auswirkungen haben Vielkommentierer (> 150 Kommentare / Monat) bzw. die Sentiments deren Kommentare im Vergleich zum Sentimen des Artikels
- Ob ein Sentiment-Trend einzelner Autoren über mehrere Artikel besteht.

4 Fazit

Da die Autoren der NYT weit aus sachlicher schreiben als initial angenommen (Sentiment Range zwischen 0 und 0.25) und online überwiegend Sachlich argumentiert und diskutiert wird, konnte keine explizite Korrelation zwischen Artikeln und deren Kommentaren eruiert werden.

Wir gehen davon aus, dass ein Boulevardmedium eine spannendere Datenquelle für eine Sentiment-Analyse gewesen wäre als die NYT.