Automatische Medieninhaltsanalyse – ein Bericht aus der Werkstatt

Bruno Wüest, Sarah Bütikofer
19th September 2016

Die Schweizer Medien berichteten während des Wahlkampfs 2015 vor allem über gesellschaftspolitische Themen und den Wahlkampf selbst. Das zeigte die Selects Medienanalyse 2015. Im Unterschied zu traditionellen Medieninhaltsanalysen hat die Selects Medienanalyse 2015 erstmals auf ein automatisiertes Vorgehen gesetzt, welches wir im Folgenden erklären. Das angewandte Verfahren der Selects Medienanalyse 2015 kann allgemein für die Bearbeitung grosser Textmengen eingesetzt werden.

Das Hauptziel der Selects Medienanalyse 2015 bestand darin, ein möglichst umfassendes Bild des Wahlkampfs in den Schweizer Print- und Onlinemedien zu erhalten. Uns interessierten dabei vor allem die Akteure (Parteien und Personen) sowie die politischen Inhalte in der Berichterstattung. Insgesamt haben wir über 275’000 Dokumente aus 93 Zeitungen, Onlineportalen und Zeitschriften in die Analyse mit einbezogen. Damit diese Textmengen präzise verarbeitet werden können, braucht es automatisierte Textanalyseverfahren. Unser Verfahren basiert auf fünf grundlegenden Schritten, die wir im Folgenden näher vorstellen.

1. Den Medienkorpus für die Wahlkampfanalyse bestimmen

Die Selects Medienanalyse 2015 basiert auf einer Zusammenarbeit mit der

Schweizerischen Mediendatenbank (SMD), welche uns Zugang auf ihr Archiv an Mediendokumenten gewährte. In einem ersten Schritt wurden die von der SMD archivierten Titel nach folgenden Kriterien gefiltert: nur Schweizer Publikationen, Erscheinungsfrequenz mindestens einmal wöchentlich, keine Fach- oder Branchenmagazine (z.B. “Die Tierwelt”). Dies ergab ein Total 93 Titel, aus welchen wir 275’705 Artikel analysierten, die sich wie folgt auf verschiedene Medienerzeugnisse verteilen:

Abbildung 1:

Überblick über den Medienkorpus

Zeitungstyp	Anzahl verfügbare Dokumente
Regionalzeitungen	93'193
Überregionalzeitungen	76'925
Newsplattformen	47'832
Gratiszeitungen	20'234
Wirtschaftsinformationen	18'300
Sonntagszeitungen	9'926
Wochenzeitungen	4'770
Illustrierte	2'812
Branchenmagazine	1'713
Total	275'705

DeFacto
BELEGT, WAS ANDERE MEINEN

Quelle: Selects Medienanalyse 2015

INFOBOX: Korpus

Eine erste Hürde ist, eine solche grosse Anzahl Dokumente effizient zu beziehen und zu speichern. Zu diesem Zweck wurde eine Pipeline von eigens erstellten Software-Skripten in die von der SMD bereit gestellte Plattform [1] eingebettet. Damit wurden Abfragen durchgeführt, Angaben zu den Dokumenten wie Publikationsdatum und Publikationstitel extrahieret und die Textdaten in Datenbanken abgespeichert. Eine effiziente Speicherung eines grossen Korpus’ ist die Voraussetzung für die nachfolgenden Analysen und Qualitätstests, welche oft wiederholt und getestet werden müssen, bis alle Berechnungen sitzen.

2. Relevante Dokumente herausfiltern

Print- und Onlinemedien berichten über ein breites Spektrum an Themen. Aus diesem Grund ist es notwendig, den Korpus zunächst nach den relevanten Dokumenten zu filtern, in unserem Fall nach den Beiträgen über den Wahlkampf. Weil die Medientitel verschiedene Rubriken – oder im Fall von vielen Online-Quellen gar keine – aufweisen, haben wir hierfür mit einer binären Klassifikation gearbeitet. Das heisst, es wurde ein statistischer Algorithmus angewendet, der aufgrund der Wortvorkommen zwischen relevanten und nicht relevanten Dokumenten unterscheidet. Ein Filter für Dokumente zur *Schweizer Politik* funktionierte von mehreren geprüften Varianten schliesslich am besten.

Diese binäre Klassifikation haben wir "überwacht" durchgeführt. Das bedeutet, dass wir zunächst eine kleine Stichprobe von Dokumenten manuell in die gewünschten Kategorien eingeteilt haben. Für unser Projekt waren drei ExpertInnen damit beauftragt, 1'813 deutschsprachige, 978 französischsprachige und 395 italienischsprachige Dokumente in die Kategorien *relevant/nicht relevant* einzuteilen.

Das ist das Anschauungsmaterial für das maschinelle Lernen. Für die Selects Medienanalyse 2015 wurden vier verschiedene Algorithmen [2] und sechs Einstellungen in Bezug auf die Verarbeitung der Texte (z.B. ob selten vorkommende Worte ausgeschlossen werden sollen), zwei Einstellungen in Bezug auf die Wortverteilungen (z.B. ob Worte nach ihrem Vorkommen gewichtet werden) und zehn Einstellungen in Bezug auf die einzelnen Algorithmen getestet. Alle diese Einstellungen wurden in allen Kombinationen mehrmals auf die Stichprobe angewandt, um systematisch den besten Algorithmus zu finden. Näheres dazu lässt sich im technischen Bericht finden.(LINK)

INFOBOX: Aufbau des Filters

In einer Pilotphase haben wir drei verschiedene Varianten eines solchen Filters getestet: Dokumente, die relevant sind für *Politik im Allgemeinen*, relevant für *Schweizer Politik* und relevant für den *eidgenössischen Wahlkampf*. Die Kategorie *Politik im Allgemeinen* erwies sich als deutlich zu breit, um relevante Dokumente für den Schweizer Wahlkampf zu finden. Ein eigenen Filter für den *Wahlkampf* hingegen funktionierte ebenfalls nicht gut, weil sich die drei ExpertInnen, welche in der Pilotphase die Tests gemacht haben, oft uneinig waren, welche Dokumente als relevant zu betrachten sind und welche nicht.

Nach einer intensiven Testphase haben wir uns auf folgende Definition von *Schweizer Politik* geeinigt: "Berichterstattung über Politik meint redaktionelle Dokumente, Meinungen oder Kommentare zu Kriegen und Konflikten, Angelegenheiten, die mehrere Staaten betreffen, Wahlen und Abstimmungen, Verabschiedung von Gesetzen, staatspolitische Fragen, öffentliche Reformen in den verschiedenen Politikbereichen sowie weitere Themen, welche die Politik direkt betreffen. Zusätzlich sind nur Dokumente relevant, in denen ein Schweizer Akteur (Person oder Organisation) vorkommt, die Schweiz erwähnt wird oder die Schweizer Innenpolitik Thema des Dokumentes ist." Folglich haben wir für die Analyse nur Dokumente einbezogen, die gemäss dieser Definition *Schweizer Politik* zum Thema hatten:

3. Umfassende Qualitätskontrolle

Das A und O eines automatisierten Filters sind umfassende Qualitätstests. Während der Entwicklungsphase kann die Qualität der automatischen Klassifikation mit der Einstufung von ExpertInnen verglichen werden. Wenn die automatische Klassifikation eine im Vergleich zur Handeinteilung ähnliche Qualität erzielt, kann die Entwicklung des Filters abgeschlossen werden. In unserem Fall erreichten die Übereinstimmung unter den ExpertInnen 0.93 in Bezug auf die Präzision und 0.67 in Bezug auf die Ausschöpfung des Filters.[3]

Abbildung 2:

Qualität der Identifikation relevanter Dokumente

Deutsch	Präzision	Ausschöpfung	Anzahl (N)
Irrelevante Dokumente	0.94	0.97	276
Relevante Dokumente	0.90	0.80	87
Gewichteter Durchschnitt	0.93	0.93	363

Französisch	Präzision	Ausschöpfung	Anzahl (N)
Irrelevante Dokumente	0.95	0.94	167
Relevante Dokumente	0.68	0.72	29
Gewichteter Durchschnitt	0.91	0.91	196

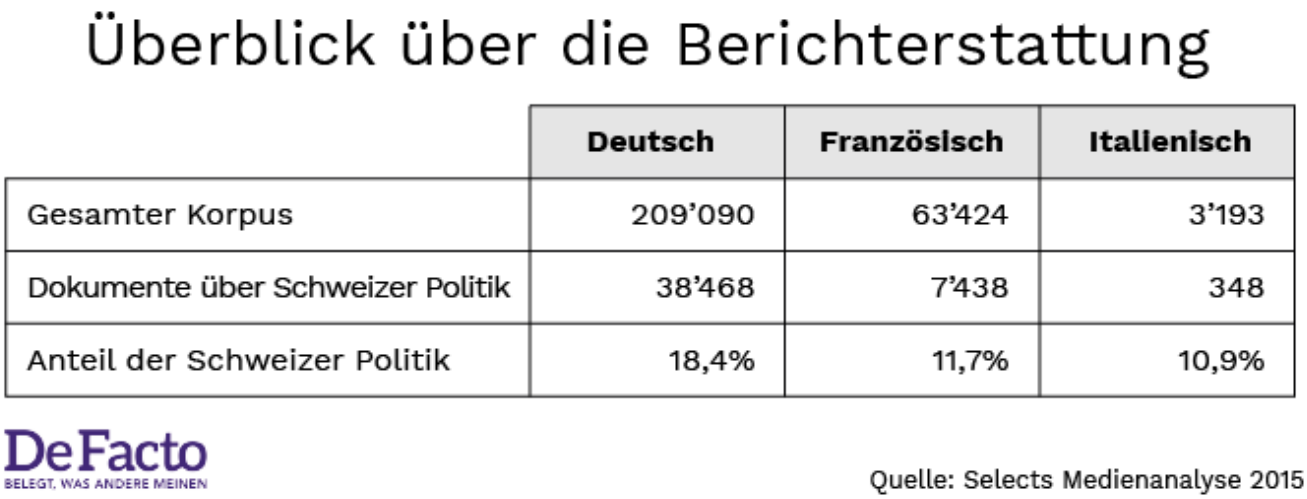
Italienisch	Präzision	Ausschöpfung	Anzahl (N)
Irrelevante Dokumente	0.88	0.91	54
Relevante Dokumente	0.78	0.72	25
Gewichteter Durchschnitt	0.85	0.85	79

DeFacto
BELEGT, WAS ANDERE MEINEN

Quelle: Selects Medienanalyse 2015

Das wichtigste Resultat unserer Analyse war, dass alle Algorithmen zufriedenstellend arbeiten. Dies, weil sie für alle Sprachen die Resultate des Vergleichs der drei verschiedenen Handcodierungen übertreffen. Darüber hinaus ist die Qualität für die Erkennung der irrelevanten Dokumente – das wäre zum Beispiel die Berichterstattung zu anderen Themen als *Schweizer Politik* – generell besser als für die relevanten Dokumente. Auch dieses Resultat ist plausibel, weil der Anteil der irrelevanten Dokumente auch höher ist.

Abbildung 3:



Als weitere Qualitätskontrolle haben wir aus den Beiträgen, die in die Kategorie *Schweizer Politik* fielen, pro Sprache je genau hundert Dokumente zufällig ausgewählt und kontrolliert. Die Präzision in diesem Test betrug 0.90 für Deutsch, 0.92 für Französisch und 0.70 für Italienisch. Das ist deutlich besser als die erste Messung für Deutsch und Französisch und gleichbleibend für Italienisch. Der italienischsprachige Korpus war besonders schwierig zu filtern, weil die Anzahl Dokumenten sowohl in den Trainingsdaten als auch im gesamten Korpus relativ niedrig ist.

4. Die Identifikation von Akteuren

Unsere erste Analyse nach der Filterung des Korpus war die Erkennung des Vorkommens von Parteien und PolitikerInnen in den relevanten Dokumenten. Hierzu haben wir zunächst zwei umfassende Listen erstellt. Für die PolitikerInnen haben wir die offiziellen Namenslisten der KandidatInnen für die eidgenössischen Wahlen 2015 mit den Namen der BundersrätInnen, ParteipräsidentInnen und abtretenden National- und StänderätInnen ergänzt. Die finale Liste umfasste 3'913 Personen.

Für die Parteienstichworte sind wir zunächst von den offiziellen Namen aller kantonalen Wahllisten für die eidgenössische Wahlen 2015 ausgegangen. Von dieser Liste wurden *Reguläre Ausdrücke* [4] gebildet, indem Duplikate gelöscht, Namen auf den Kern reduziert (z.B. "Lega" anstatt "Lega dei Ticinesi") und verschiedene Endungen sowie Gross- und Kleinbuchstaben antizipiert wurden (so dass z.B. die 'schweizerische', die 'Schweizerische' und die 'Schweizerischen Volkspartei' gefunden wird) wurden. Zudem haben wir Synonyme für Begriffe hinzugefügt (z.B. "Freisinn" für die "FDP.Die Liberalen"), welche wir aus früheren Inhaltsanalysen zur Verfügung hatten (vgl. Wueest, Müller und Willi 2016). Die Liste der Parteien umfasste schlussendlich 181 Stichworte.

Beide Listen wurden intensiv in mehreren Iterationen getestet. Während der Erstellung der Listen haben wir alle Parteienstichworte und eine Stichprobe von KandidatInnennamen auf der normalen SMD-Benutzeroberfläche getestet, d.h. die Stichworte und Namen mit hohen sowie sehr wenigen Trefferquoten systematisch auf ihre Genauigkeit überprüft. Um die Leistungsfähigkeit der

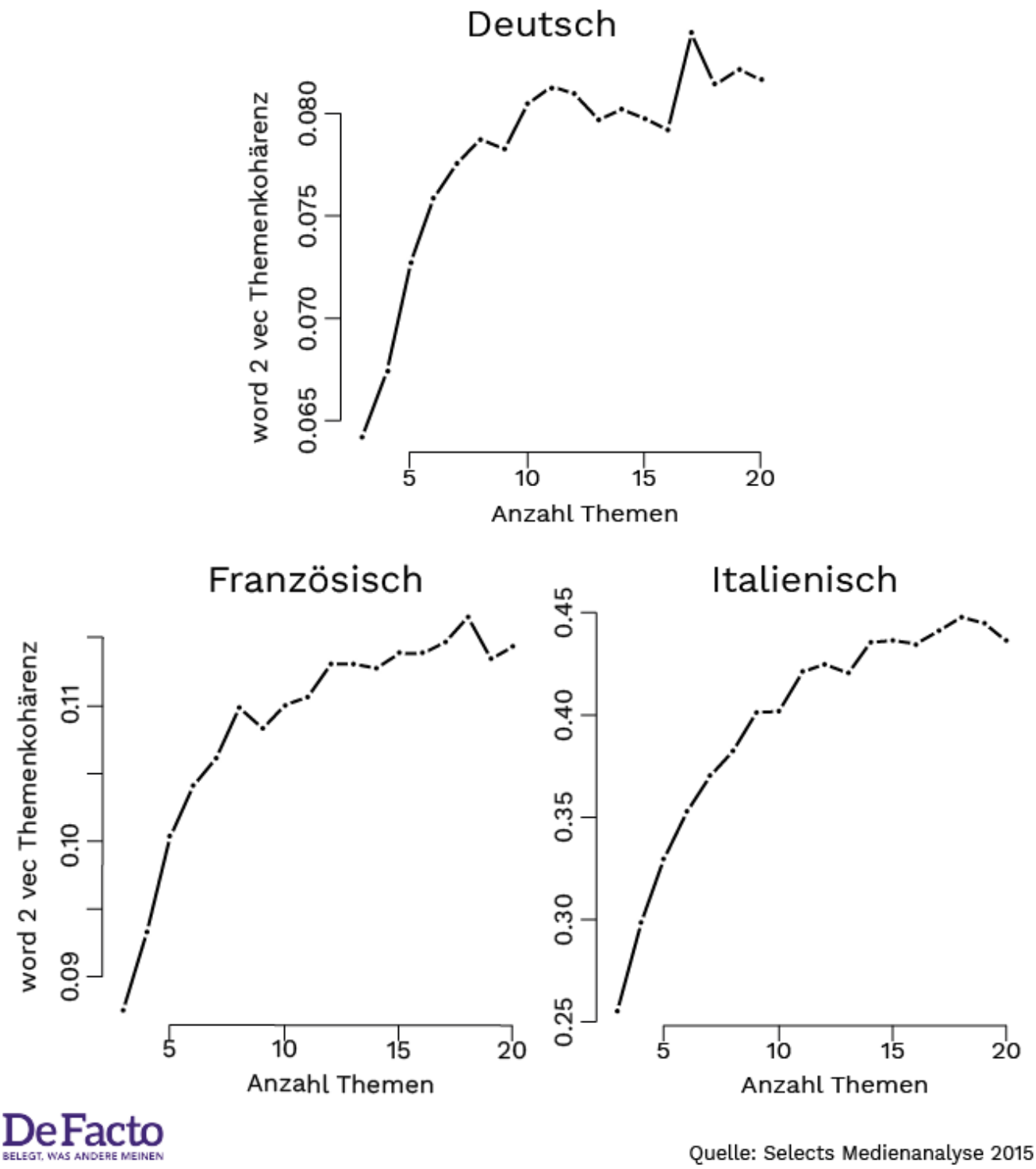
Analyse zu erhöhen, haben wir durch Parallelisierungen bis zu dreissig Suchen gleichzeitig durchgeführt, womit sich die Laufzeit auf ca. zwei Stunden reduziert hat.

5. Thematische Schwerpunkte erkennen

Um die Themen in den relevanten Dokumenten zu erkennen, haben wir uns für ein induktives Vorgehen entschieden. Die Themen, sogenannte topics, wurden direkt aus den Dokumenten berechnet, und zwar mit strukturellen Themenmodellen (Structural Topic Models, STM, Roberts et al. 2014). Die STM schätzen die Wahrscheinlichkeit, dass ein Dokument zu einem bestimmten latenten Thema gehört.[5] Die STM ergeben als Resultat Listen von Wörtern, welche für die einzelnen Themen typisch sind. Mit diesen Wortlisten und mit der Lektüre von typischen Dokumenten konnten wir schliesslich eine Einteilung der in den analysierten Beiträgen gefundenen Themen in Wahlkampfthemen vornehmen.

Die folgenreichste Entscheidung bei der Anwendung eines Themenmodells ist die Anzahl Themen, welche man vor der Berechnung angeben muss. Ein Modell mit zu wenigen Themen produziert zu diffuse Themenkategorien, wohingegen ein Modell mit zu vielen Themen zu sehr spezifischen und fast nicht unterscheidbaren Themenkategorien führt. Wir lösen dieses Problem, indem wir die richtige Anzahl Themen aufgrund der Kohärenz der Worte, welche für ein Thema wichtig sind, berechnen.[6] Die nachfolgende Abbildung zeigt, dass die Kohärenz der Worte für Italienisch und Französisch auf 18 Themen hinweist, und für Deutsch auf 17.

Abbildung 4:



Konsequenzen der technischen Neuerungen

Die Selects Medienanalyse 2015 setzte im Vergleich mit früheren Analysen auf ein konsequent automatisiertes Verfahren. Dadurch konnten zum einen die Inhalte ganzer Zeitungen und zum anderen Medientitel aus drei Sprachen und aus allen Landesteilen in die Analyse einbezogen werden.

Automatisierte Textanalysen haben aber nicht nur Vorteile. Die Arbeitslast verschiebt sich im Vergleich zu manuellen Inhaltsanalysen von der eigentlichen Datenerhebung hin zur Kontrolle der Datenerhebung. Gerade, weil die Berechnungen weitestgehend automatisiert sind, braucht es für eine präzise Analyse ein Vielfaches an Qualitätstests.

Abschliessend können wir festhalten, dass die in der Selects Medienanalyse 2015 eingesetzten Verfahren generell einsetzbar sind. Wir hoffen, dass unsere Pionierstudie als Anleitung für vergleichbare Analysen zur Schweizer Politik und Medien dient.

An dieser Stelle möchten wir uns noch einmal bei der Schweizerischen Mediendatenbank SMD für den grosszügig gewährten Zugang zu ihrem Archiv bedanken.

[1] KNIME-Anbindung (Konstanz Information Miner, https://www.knime.org/ an das Apache Solr/Lucene interface des SMD.

[2] Support Vector Machine, Naïve Bayes, Random Forest und Kernel Ridge Regression

[3] Die Präzision eines Vergleichs gibt an, wie viele der als relevant eingestuften Dokumente tatsächlich relevant sind. Die Ausschöpfung gibt an, wie viele aller relevanten Dokumente auch tatsächlich als relevant eingestuft werden.

[4] Reguläre Ausdrücke sind Zeichenketten, welche bestimmte Sprachmuster abbilden und somit zu generellen Suchbegriffen ausgebaut werden können. Eine einfache Anwendung ist der Asterisk in Google-Suchen, der als Platzhalter für irgendein unbekanntes Wort eingesetzt werden kann.

[5] Die STM basieren auf der Latent Dirichlet Allocation, einem generativen Wahrscheinlichkeitsmodell, mit dem die Zugehörigkeit von Dokumenten und Worten zu den Themen geschätzt wird. Näheres dazu lässt sich im technischen Bericht finden.

[6] Wir verwenden *word2vec* zu diesem Zweck. Näheres dazu lässt sich im technischen Bericht finden.

Referenzen:

- Wüest, Bruno, Sarah Bütikofer, Fionn Gantenbein, Adrian van der Lek (2016). *Selects Media Analyses 2015. Codebook and Technical Report*. Zürich: IPZ.

- Wüest, Bruno, Christian Müller und Thomas Willi (2016). Exploring the usefulness of Twitter data for political analysis in Switzerland. Paper presented at the Annual Conference of the Swiss Political Science Association at the University of Basel, January 21-22, 2016.

Titelbild: Pixabay

Grafiken: Pascal Burkhard