Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges - *by J. Wilkerson & A. Casas

1. Introduction

Revolution für Schaffnung von Forschungsmöglichkeiten durch das Internet, welches große Menge an politischen Daten bereitzustellen vermag

- z.B. digitale Bereitstellung von Regierunsunterlagen, online Archive von Zeitungen etc.
- Antwort auf große Menge durch Aufbau von open-source Textanalyse libraries in R, Python & anderen Programmiersprachen

2. Four Stages of a Text as Data Project

- Methode erweitert Forschungsmöglichkeiten auf zwei Wegen

1. Nutzen Rechenleistung um ambitionierte Datenkollektionen möglich zu machen
2. Bieten wachsende Zahl an Optionen zur Analyse großer "volumes" quantitativer texte

- typischerweise beinhaltet ein solches Project 4 Phasen: 1. Text muss beschafft/obtained werden 2. konvertiert zu quantitativen Daten 3. analysiert und 4. validiiert werden

2.1 Obtaining Text

- Download digitalisierter Inhalte
- Unterschiede zwischen Quellen im Bezug auf Einfachheit der Infobeschaffung
- API: Ermöglichen es Nutzern ausgewählte Inhalte anzufragen; APIs sind ideal, wenn sie Optionen die den Bedarf des Projekts abdecken, beinhalten
- Falls API nicht vorhanden ist, Rückgriff auf Dokumente mit ähnlicher Formatierung -> identische Formatierung macht es möglich ein Skript zu schreiben, dass spezifische Inhalte von vielen Dokumenten gleichzeitig extrahieren kann.
- beinahe alle Dokumente besitzen versteckte Formatierungssprachen, die nützlich für Extraktion systematischen Inhalte sein können
    - Aussehen einer Webseite entsteht durch eingebettetet HTML oder XML Tags -> Tags können genutzt werden um gewünschten Inhalt zu isolieren
- Schwiergisten Text Extraktions ("Scraping") Projekte sind jene die Inhalt von verschiedenen Quellen beziehen
    - z.B. Extraktion selben Inhalts von verschiedenen Kandidatenwebseiten, weil jede Webseite seine eigene Struktur besitzt
        - ♠ Option: 1. Multiple Skripts verfassen -> siehe OpenStates project; 2. Einfachere Metriks sammeln z.B. Anzahl von Keywörtern -> gewöhnlicher Ansatz bei "big data" projects

2.2 From Text to Data

Inhalt jeden Dokuments muss in quantitative Daten umgewandelt werden

- manchmal Zielsetzung eine *term-document* oder *term-frequency* Matrix zu erstellen, in der jede Reihe ein Dokument und jede Spalte ein gefundenes Feature aus einem der Dokumente darstellt
- in dieser Phase muss Forscher sich für eine passende Analyseeinheit entscheiden
- im nächsten Schritt wird spezifiziert welches Feature innerhalb eines jeden Dokuments in der quantitativen Analyse genutzt wird
    - Startpunkt: jedes einzigartige Wort als ein seperates Feature behandeln
    - Forscher entfernen Dokumentinhalte die als irrelevant und potentiell irreführend für die Analyse angesehen werden
        - Standardoptionen hierfür: Punktuation entfernen; common words (stopwords), very infrequent words (sparse terms) & word suffixes (stemming)
- im nächsten Schritt Features erstellen die über basic Worte hinausgehen
    - Wortpaare einsetzen (bigrams) als zusätzliche Features
    - Anstatt Synonyme als separate Worte anzusehen kann man diese als einzelnes Feature zusammenfassen
    - wichtigeren Features höhreren Wert zuweisen

2.3 Quantitative Analysis of Text

Einfache Metriken können nützlich sein und haben zusätzlichen Vorteil der Transparenz & Replizierbarkeit

- z.B. Nutzt Casas Liste von positiven & negativen Worten um zu studieren, wie die Medien Protestierer darstellen
- heutzutage Fokus auf statistischen machine learning Methoden
- Diskussion über Unterschied von machine learning & Statistik
    - Politikwissenschaftler nutzen statistische Methoden um Theorien zu testen -> wählen bestes Modell für die Daten (ordinary least squares, logistic regression etc.) bevor Modellspezifikation getesten werden, die eine begrenzte Anzahl von theoretiisch hergeleiteten Input (independent) Variablen beinhalten
        - Fokus auf Koeffizienten oder Parametern für die Input Variablen -> z.B. Frauen signifikant wahrscheinlicher Demokraten als Männer
        - Ob das Model die Parteiidentifikation eines jeden Wählers richtig vorhersieht ist sekundäres Bedenken
    - Bei machine learning Untersuchungen liegt der Fokus gewöhnlich auf dem Output, als dem Input
        - anstatt zu Fragen ob Frauen Demokraten sind, liegt Zielsetzung bspw. darin die politische öffentliche Meinung auf Landesebene zu prognostizieren
        - Fokus auf Output führt dazu, dass Forscher sich mehr mit der Genauigkeit ihrer Prognose beschäftigen & weniger mit dessen Begründung
        - Fokus auf Prognostizierung fördert ebenso die Experimentierfreudigkeit mit verschiedenen Algorithmen & Features

2.4 Evaluating Performance

Validiierung ist wichtige Komponente eines jeden Text-as-data projects

- beaufsichtigte machine learning Ergebnisse werden durch den Vergleich der Prognose eines Algorithmus zu den vor exitierenden "gold standard" Ergebnissen validiiert
    - "gold standard" für Beauchamp: sind Landesweite Meinungsumfragen
- In Informatik nutzen Forscher online Bewertungen & Reviews um Algortihmen welche Sentimente festhalten zu trainieren & validiieren
    - um es vor Überanpassung zu schützen, wird der Algortihmus zunächst mit einem Set von beschrifteten/Etikett(labeled) Beispielen trainiert, anschließend wird die Genauigkeit mit einem anderen "held-out Set" getestet
- andere Methoden bei denen Gold Standard nicht möglich ist, findet die Validiierung vielfältig statt

    - unüberwachte machine learning methoden: Forschungen vertiefen sich in spezifische Beispiele innerhalb des Themas um zu zeigen, dass das Thema sinn ergibt, um zu zeigen das andere Algortihmen ähnliche cluster erzeugen

3. Recent Developments in Political Science

Untersuchungsmöglichkeiten dargelegt. 4 generelle Untersuchungsziele -> Klassifikation, Scaling, Text reuse & semantics

3.1 Classification

Nicht beaufsichtigte machine learning Methoden (z.B. PCA, latent Dirichlet allocation LDA) vergleichen die Ähnlichkeit von Dokumenten basierend auf den vorkommenden Features

- benötigen viel Imput vom Nutzer, welcher die Anzahl von Topics im vorhinein spezifisieren & ihre Bedeutung interpretieren muss

Wo unsupervised Methoden oft zur Entdeckung genutzt werden, werden supervised learning Methoden primär als Arbeit sparendes Mittel genutzt

- Tatsache das supervised Methode oft mehrere tausende von Traningsbeispielen benötigt, macht es für viele Forscher zum nonstarter
- Wege um benötigten Effort zu reduzieren: Ausschluss von Textinhalten durch einfache Identifikationstechniken
- Crowdsourcing oft genutzt um Übungssets in computer science zu bilden
- wenn ein Projekt keine individuellen Dokument Etiketts braucht, ist ReadMe eine supervised Methode die reliabel Klassenproportionen prognostiziert, indem es eine kleinere Zahl von Übungsbeispielen braucht

Sentiment Analyse ein weiterer wichtiger Bereich der Klassifikations Untersuchungen in der (un-)supervised Methoden genutzt werden

- Ziel ist es den Text ordinal zu klassifizieren (von negativ zu positiv z.B.), anstatt es kategorisch vorzunehmen
- Sentiment Analyse gut finanziert, daher viele vorexistierende Tranings corpora für eine weite Anzahl von Untersuchungsgebiete vorhanden

3.2 Scaling

Eine der frühesten Applikationen von automatisierter Text Analyse fokusierte sich auf die Nutzung von Reden & Manifestos um europäische politische Parteien im ideologischen Raum zu lokalisieren

- spätere Forschung erweiterte dies durch die Verwendung neuer Methoden & Untersuchung neuer Bereiche
- Benoit: zeigte das crowdsourcing eine mögliche Alternative zu expertenbasierten Ansätzen sein kann, um Parteien auf politischen Dimension zu lokalisieren
- Kluver: nutzte Statements von Interessengruppen & EU regulators
- Barbera: Nutzte Twitter Daten & Infos über posters' followers um die ideologische Position von Politikern, Parteien & Bürgern zu bestimmen

3.3 Text reuse

Impliziert die Entdeckung von Fällen von ähnlichem Sprachgebrauch

- Unterscheidungsmerkmal des Text reuse Algorithmen ist dass sie die Wortfolge bei der Beurteilung der Dokumentähnlichkeit explizit bewerten
    - aktuell genutzt um Ursprünge von politischen Vorschlägen in der Gesetzgebung zu verfolgen, um den Einfluss von Interessengruppen auf den Gesetzgeber zu untersuchen
- andere Möglichkeiten: Untersuchung der Diffusion von politischen memes & dem ansteckenden Effekt in neuen und alten Medien
- verschiedene Algorithmen unterstützen verschiedene Analysetypen
    - global ausgerichtete Ansätze, identifizieren & bewerten die geteilten Wortsequenzen innerhalb von Dokumenten

Natural Language Processing

Analyse von sozialen Netzwerken verwendet Text um die Beziehung zwischen Akteuren zu beobachten

- Natural language processing (NLP) macht es möglich geht weiter als simple Verbindungen zu etablieren, sondern den Status der Beziehung zu untersuchen -> moving from "whomw" to "who did what to whom"
- anstatt die Anzahl von Nennungen zweier Akteure in einem Bericht zu zählen, event data analysis integrieret sie Syntax und Semantiks um systematisch zu beobachten ob die Beziehung besser oder schlechter wird
- frühe Formen von event data Untersuchungen waren abhängig von menschlichen Annotators, damit diese dictionaries von Objekten & Handlungen zu entwickeln
- aktuelle Forschung zielt darauf die Blickweite dieser Untersuchungen auszuweiten, indem Vorteil von extensiven NLP Ressourcen genommen wird
- Andere Ressourcen: Wortnet -> kann genutzt werden um Synonyme für ähnliche Handlungen oder Sentimente zu finden

4. Topicmodel instability and a call for greater attention to robustness in text-as-data research

Unsupervised machine learning Mehoden (topic models) sind popülär in Politikwissenschaft, teilweise weil es Dokumente ohne extensive Etteketierungsmühen, wie es bei supervised learning methods nötig ist, klassifiziert

- für gewöhnlich wird zunächst ein einziges topic model gemeldet & validiiert, nachdem Ergebnisse diverser Modelle die sich durch ihre Anzahl von topics spezifisiziert durch den Forscher unterscheiden verglichen wurden
    - Wahl basiert auf Forschers subjektiver Einschätzung darüber welches model cluster die wesentlichen Ziele des projekts am besten reflektiert
- Erschwerung der Validiierung (1. Fehlender gold standard): Als zweites Problem tritt die Model Instabilität auf
    - Chuang: Bestimmte die selben 50 topic models, um zu sehen, dass nur 2 von 25 topics über alle BEstimmungen hinweg, beständig waren
        - ♠ kann passieren, weil verschiedene Einschätzungen auf verschiedenen lokalen Maxima konvergieren
    - Manipulation eines FEatures eines strukruellen topic models kann zu anderen Ergebnissen führen
- Antwort auf Modelinstabilität: Auswahl und Validiierung eines best models
    - bei konventionellen statistischen Studien, versuchen Forscher zu demonstrieren, dass ihre Ergebnisse robust sind durch die Darlegung von Ergebnisse für multiple Modellspezifikation
    - supervised machine learning Analysen adressieren Robustheit dadurch das Ergebnisse auf basis der Konsens Prognostizierung eines Ensembles darstellen, anstelle eines einzelnen Algorithmus

4.1 Exploring the Topics of Legislators' Floor Speeches REST NUR BEISPIEL

Darstellung wie topic robustness eine Studie of congressional floor speeches informieren kann

- US House of Representatives ca. 10.000 "one-minute" Ansprachen während des 113ten Kongress 2013-2014 -> Reden vor ordinary business und primär für öffentliche consumption beabsichtigt
- Überblick indiziert, dass Subjekte oft sehr unterschiedlich sind
- Nutzung des Sunlight Foundation's Capitol Words API um die Statements aller Mitglieder vom Congressional Record runterzuladen
- Statements wurden entfernt, welche nicht mit opening phrase einer one minute speech "Mr. Speaker, I rise today..." anfingen -> Ergebnis corpus von 5.346 one minute speeches von 179 Demokraten und 4.358 von 213 Republikanern
- alle Worte in lower case konvertiert & Punktuation entfernt, sowie stopwords, word stem & Wörter mit 2 oder weniger Charakteren
- zuletzt term-document matrix erstellt in der jede Reiehe eine one minute speech und jede Spalte ein Vektor indizierend ob ein Feature/Wort in einer Rede präsent war
- nächster Schritt Schätzung eines latent Dirichlet allocation LDA models wo die Anzahl von topics (k) von 10 bis 90 reicht und in 5 topics increments
- 17 Modelle führen zu 850 topics -> um zu bestimmen welches topic robust ist wird cosinus similarity für alle topic paare und anschließend der Spectral Clustering algorithmus genutzt um die 850 auf basis der cosinus similarity zu gruppieren
- Spectral Clustering Algorithmus tut dies durch Maximierung der durchschnittlichen intra-cluster cosinus similarity für eine gegebene Anzahl von clusters c
- Gehalt eines gegegeben CLusters kann dann durch Untersuchung der meist prognostizierten worte (top terms) in jedem cluster investigiert werden
- Figure 2 stellt dar wie die Variation der Anzahl von speech clusters (c) zwischen 5-100 die Passung verbessert
    - mehr Cluster verbessern die allgemeine Eignung bis ca 50 cluster
    - Analyse basiert daher auf robusten topics von 50- cluster Modelen
- Figure 3: Nach dem Clustern der 850 topics von 17 Modelen in 50 cluster, werden einige cluster gruppiert in metatopics
- Figure 4: z.B. beinhaltet metatopic Bildung 37 topics aus 14 verschiedenen topic models
    - dieses Figure unterstreicht den Nachteil Ergebnisse auf Basis eines einzigen Models zu zeigen - Topics that are common to many models are often missing from any one of them

4.2 Topic Attention in One-Minute Speeches

In einem LDA model existieren die topics vor den Dokumenten. Es wird angenommen, dass jedes Dokument mit positiver Wahrscheinlichkeit über jedes topic ist.

To study speech attention, we must first label individual speeches for primary topic. We assume each speech is about its most probable topic.

Thus, we classify 9,704 speeches for each of 16 topic models. We then report results for only those topics from each model that are part of the 21 metatopics. Figure 5 displays those results.

4.3 Validation

Similarity of estimates of topic emphasis across different models is an important type of validation

Figure 5 should inspire confidence in the robustness of general differences in speech topic emphasis but less confidence in the amount of difference in many cases

5. Discussion

Über Fortschritt den computergestützte Text Analyse für die politische Wissenschaften bietet