Research Methods in the Age of Digital Journalism

Massive-scale automated analysis of news-content

Hendrik Heuer  
hendrik.heuer@aalto.fi  
Special Course in Language Technology  
Aalto University

Introduction

• **Traditional social science**: required **coding teams working for several months** to produce the data

• combination of **computer science and social sciences** makes **new research** possible
  - analysis of massive social networks
  - content of millions of books
  - data mining, machine learning and natural language processing
Study

- large-scale investigation

- **2.5 million articles** from **498** different English-language **news outlets**
  
  *Reuters* and *New York Times Corpus*

- automatically annotated into 15 topic areas

- the topics were compared in regards to **readability**, **linguistic subjectivity** and **gender imbalances**
Method

- **Discovery of adjectives:**
  Stanford Log-linear Part-of-Speech Tagger

- **Gender of people:**
  Open-Source Tool Gate

- **Readability:**
  Flesch Reading Ease Test
  (based on word and sentence length)

- one **SVM classifier per topic**
Support Vector Machine (SVM)

- Machine Learning technique
- Choose the separation that has maximal margins
- Transform input data non-linearly into a high-dimensional feature space
Findings

• as we might expect:
  • Op / Ed pieces more linguistically subjective
  • News for children are more readable
  • Articles about female sports are more likely to mention women
Readability

Comparison of topics based on readability
Readability
Comparison of topics based on readability
Linguistic subjectivity

Comparison of topics based on linguistic subjectivity
Linguistic subjectivity

Comparison of topics based on linguistic subjectivity
“The more readable a topic is, the more linguistically subjective it tends to be”
Writing Style

Comparison of topics based on their writing style
Male / female ratio

Comparison of topics based on their male to female ratio
Male / female ratio

Comparison of topics based on their male to female ratio

“almost equal references to males and females”
“Gender bias in sports coverage (...) females only account for between only 7 and 25 per cent of coverage”
“Gender bias in sports coverage (...) females only account for between only 7 and 25 per cent of coverage”

The sex ratio for the entire world population is 101 males to 100 females.

Source: Sex Ratio http://en.wikipedia.org/wiki/Human_sex_ratio
Writing Style

Comparison of a selection of newspapers from the USA and the UK based on their writing style
Writing Style

Comparison of a selection of newspapers from the USA and the UK based on their writing style
Writing Style

Comparison of a selection of newspapers from the USA and the UK based on their writing style
Topics

Comparison of a selection of newspapers outlets based on the topics they choose to cover
Topics

based on a 15-dimensional vector, projected using Multidimensional Scaling
Popular articles tend to be more readable and more linguistically subjective.
Conclusion

• **Failure** of many **citizens to engage with political news or environmental problems** like climate change may reflect the fact that **these news topics** tend to be **less readable**

• **Low levels of political interest** and **engagement** might be connected to the **lack of** adjectival excess or **subjectivity**
Conclusion

• Findings confirm features of media content that have been predicted on the basis of theoretical assumptions and smaller-scale content studies

• Allows social sciences to be both more ambitious and more comprehensive in scale
Conclusion

• Automated news content analysis will not replace human judgement, but it allows much bigger sample size and frees scholars from mundane tasks.

• Allows to focus on higher level properties and account for the vast, multi-dimensional communications systems we have (not only newspapers and TV).
Comments
“To measure the **popularity of articles** we tracked the special news feed provided by some outlets that carries the “**Most Popular**” articles.”

**Comparability problem,** the Guardian might evaluate this very differently from the New York Times and the Daily Mail.
“Absence of any gold standard for sentiment analysis” (Pang and Lee 2008)

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org, {aperelyg, jcchuang, ang}@cs.stanford.edu
{jeaneis, manning, cgpotts}@stanford.edu

Abstract

Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network. When

Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (−−−−−, 0, +, ++), at every node of a parse tree and capturing the negation and its scope in this way.
Computational Social Science

David Lazer,† Alex Pentland,‡ Lada Adamic,§ Sinan Aral,∥ Albert-László Barabási,¶ Devon Brewer,* Nicholas Christakis,† Noshir Contractor,∥ James Fowler,∥ Myron Gutmann,∥ Tony Jebara,§ Gary King,† Michael Macy,∥∥ Deb Roy,∥ Marshall Van Alstyne,*

We live in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based on an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the
Computational Social Science

David Lazer,1 Alex Pentland,2 Lada Adamic,3 Sinan Aral,2,4 Albert-Laszlo Barabasi,5 Devon Brewer,6 Nicholas Christakis,1 Noshir Contractor,7 James Fowler,8 Myron Gutmann,3 Tony Jebara,9 Gary King,1 Michael Macy,10 Deb Roy,2 Marshall Van Alstyne1,11

We live in a network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors. Either scenario will interest of accumulating knowledge. Computational social science could become the exclusive domain of private companies and government agencies from which they produce academic environments understanding understandings? What are the emerging field of data-driven “computational social science”

could become the exclusive domain of private companies and government agencies
Research Methods in the Age of Digital Journalism

Massive-scale automated analysis of news-content

Hendrik Heuer
hendrik.heuer@aalto.fi
Special Course in Language Technology
Aalto University