gesis Leibniz Institute for the Social Sciences

Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

GESIS workshop - November 2021

Leibniz Association

gesis Leibniz Institute for the Social Sciences

Welcome to Day 2!

Recap

- We started with a general introduction of working with social media data —
  - some of its potentials in social science research but also some limitations
- we saw various ways of collecting data, mainly through APIs
- we saw an example of collecting data programmatically from Reddit

Today's Schedule

Tuesday, 06.12.	
9:30-11:00	Recap from yesterdays activities. Time for Questions. Introduction to Preprocessing and Analysis
11:00-11:15	*Break*
11:15-12:00	Exercise for Data Preprocessing and analysis
12:00-12:30	*Break*
12:30-14:00	Introduction to existing frameworks for Error Identification and Characterization

gesis Leibniz Institute for the Social Sciences

2. Data (pre)processing and analysis

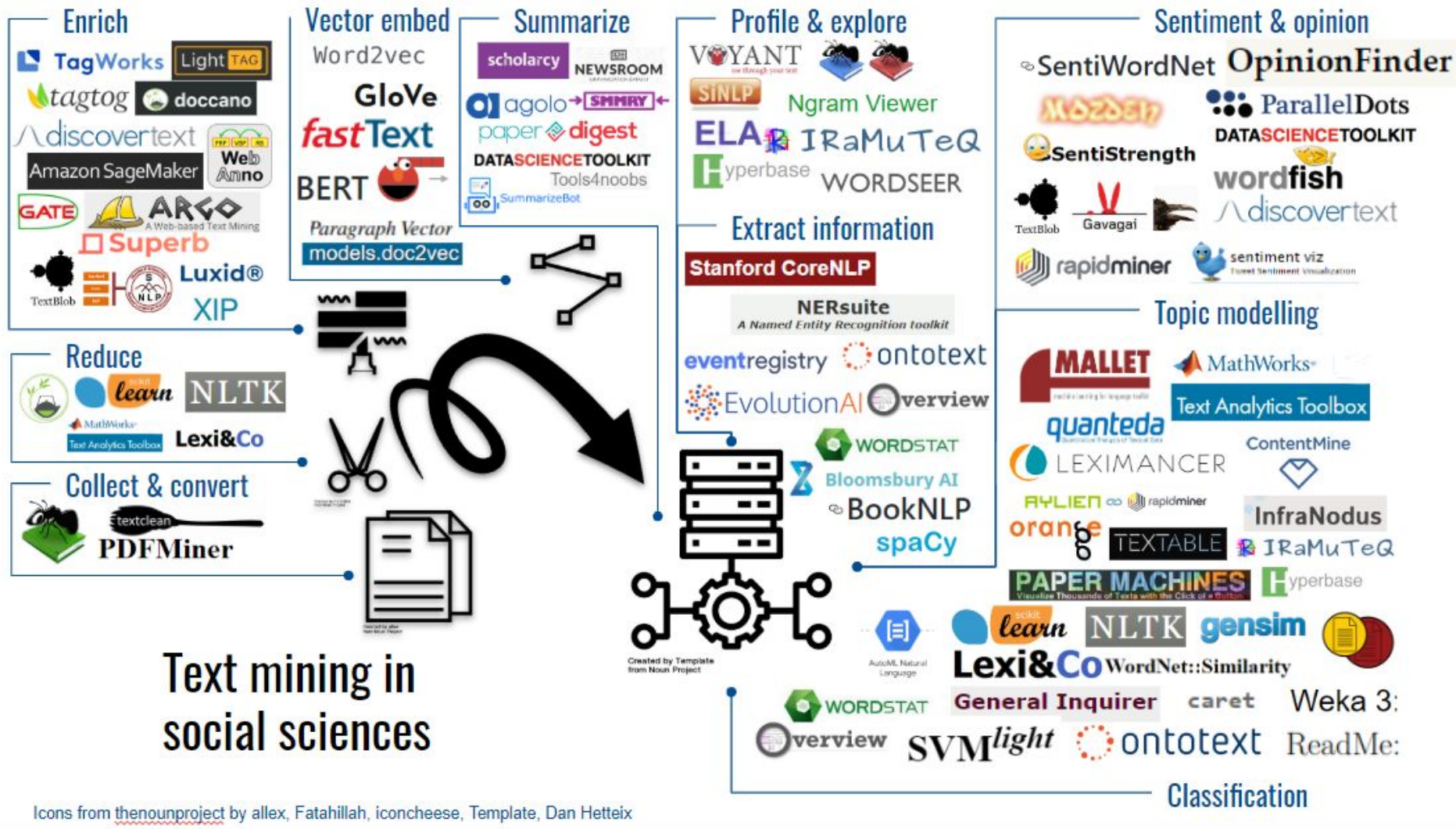
Photo: Katrin Weller

Various ways to analyze data

- our focus is on textual data
  - e.g. topic detection, sentiments
- but in general much more options related to
  - image analysis, multimedia data
  - network analyses
  - timelines
  - and more

Data Preprocessing and Analysis

- sometimes, there isn't a clear separation between the two steps and preprocessing and analysis may go hand in hand
- steps may range from text cleaning (parsing the data, removing stopwords) to complex analysis (create text networks or get stance of the post)



Icons from thenounproject by allex, Fatahillah, iconcheese, Template, Dan Hetteix

Taken from PACSS 2020 Sage Ocean Talk

Importance of Data Cleaning

Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It*

Matthew J. Denny† Arthur Spirling‡

Abstract

Despite the popularity of unsupervised techniques for political science text-as-data research, the importance and implications of preprocessing decisions in this domain have received scant systematic attention. Yet, as we show, such decisions have profound effects on the results of real models for real data. We argue that substantive theory is typically too vague to be of use for feature selection, and that the supervised literature is not necessarily a helpful source of advice. To aid researchers working in unsupervised settings, we introduce a statistical procedure that examines the sensitivity of findings under alternate preprocessing regimes. This approach complements a

Importance of Data Cleaning

Text Preprocessing For Unsupervised Learning: Why It Matters,

Despite
research, th
have receiv
found effect
theory is ty
literature is
unsupervise
ity of findi

ELSEVIER

Information Processing & Management

Volume 50, Issue 1, January 2014, Pages 104-112

The impact of preprocessing on text classification

Alper Kursat Uysal, Serkan Gunal

Show more

https://doi.org/10.1016/j.ipm.2013.08.006

Importance of Data Cleaning

1. Remove unnecessary items from our dataset
   a. function words (‘the’, ‘on’, etc)
2. Maintain order and consistency.
3. Standardization.

Data Cleaning: Options

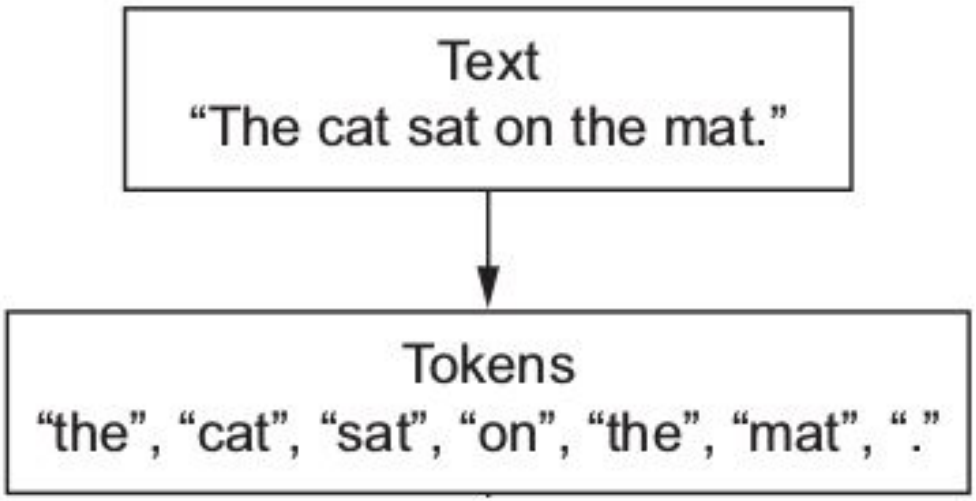
- click and point
  - PDFMiner
  - coming soon: Sage Ocean's Texti: https://ocean.sagepub.com/texti
- programmatic:
  - part of many NLP and ML libraries:
    - sklearn, spacy, **nltk**, huggingface, **pandas**

Data Cleaning: Hands On

- upload Reddit data we just collected
- Do the following steps:
  - tokenization
  - remove stopwords
  - Stemming / Lemmatization

Data Cleaning: Hands On

- upload Reddit data we just collected
- Do the following steps:
  - tokenization
  - remove stopwords
  - augm
  - augm
  - [Try it



Tokenization of a sentence

Data Cleaning: Hands On

- upload Reddit data we just collected
- Do the following steps:
  - tokenization
  - remove stopwords
- Stemming / Lemmatization

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Data Cleaning: Hands On

- uplo
- Do t
  - t
  - r
  - S

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Examples of stemming

Data Cleaning: Typical Steps

- tokenization
- remove stopwords
- Stemming / Lemmatization
- remove numbers
- remove headers and footers
- remove rare words
- Beyond Text:
  - dropping or imputing missing values
  - dropping columns with missing values

Let's try it!

Data Analysis: Hands On

- Which candidate has the most positive/negative posts?
- How to aggregate into a single metric?
- Bonus: which candidate is the most insulted?

Data Analysis: Hands On

- Augment with sentiment
- Augment with toxicity
- Topic modeling
- [Try it yourself] augment with entities

Data Analysis: Sentiment Analysis

Data Analysis: Sentiment Analysis

Sentiment Analysis is the task of automatically annotating the sentiment / polarity of a piece of content.

- Traditionally lexica-based
- Traditionally methods were built using reviews of movies or items
- Currently several sophisticated **Machine Learning** Methods exist

Data Analysis: Sentiment Analysis

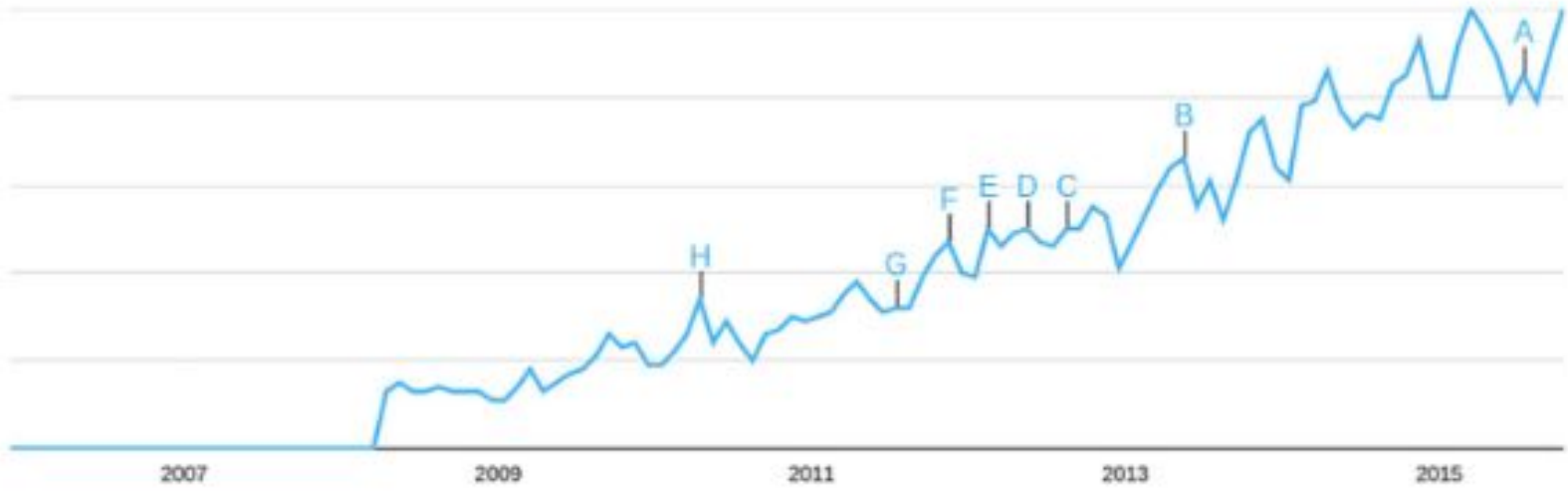


Figure 1 Searches on Google for the Query: 'Sentiment Analysis'. This figure shows the steady growth on the number of searches on the topic, according to Google Trends, mainly after the popularization of online social networks (OSNs).

Data Analysis: Sentiment Analysis

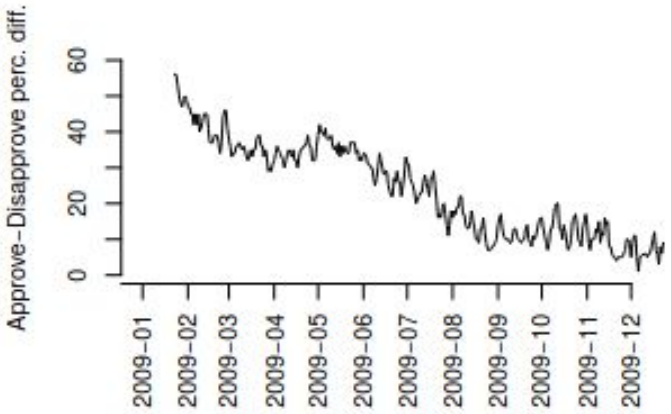


Figure 2: 2009 presidential job approval (Barack Obama).

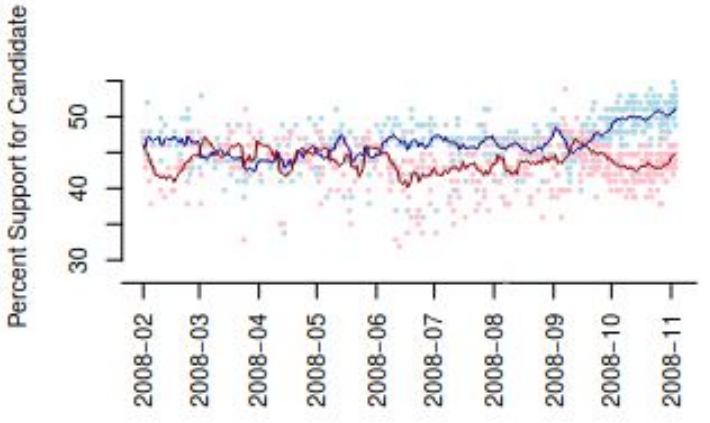


Figure 3: 2008 presidential elections, Obama vs. McCain (blue and red). Each poll provides separate Obama and McCain percentages (one blue and one red point); lines are 7-day rolling averages.

Data Analysis: Sentiment Analysis

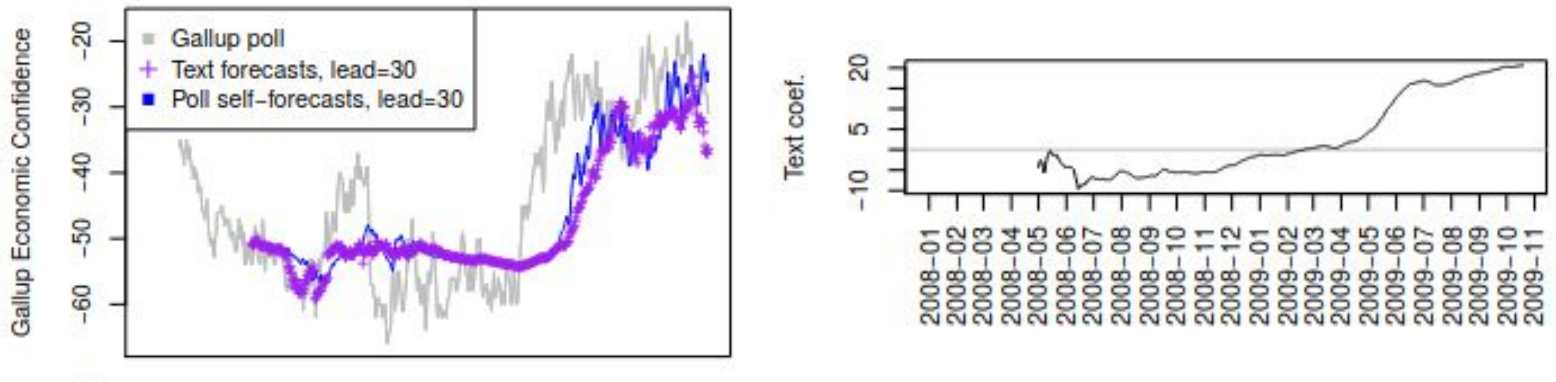


Figure 8: Rolling text-based forecasts (above), and the text sentiment (MA_t) coefficients a for each of the text forecasting models over time (below).

Data Analysis: Sentiment Analysis

Ribeiro et al. *EPJ Data Science* (2016) 5:23
DOI 10.1140/epjds/s13688-016-0085-1
EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods

Filipe N Ribeiro[1,2*], Matheus Araújo[1], Pollyanna Gonçalves[1], Marcos André Gonçalves[1] and Fabrício Benevenuto[1]

*Correspondence:
filiperibeiro@dcc.ufmg.br
[1]Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Brazil
[2]Computer and Systems Department, Federal University of Ouro Preto, Joao Monlevade, Brazil

Abstract

In the last few years thousands of scientific papers have investigated sentiment analysis, several startups that measure opinions on real data have emerged and a number of innovative products related to this theme have been developed. There are multiple methods for measuring sentiments, including lexical-based and supervised machine learning methods. Despite the vast interest on the theme and wide popularity of some methods, it is unclear which one is better for identifying the polarity (i.e., positive or negative) of a message. Accordingly, there is a strong need to conduct a thorough apple-to-apple comparison of sentiment analysis methods, *as they are used in practice*, across multiple datasets originated from different data

Data Analysis: Sentiment Analysis

Table 1 Overview of the sentence-level methods available in the literature

Name	Description	L	ML
Emoticons [20]	Messages containing positive/negative emoticons are positive/negative. Messages without emoticons are not classified.	✓	
Opinion Lexicon [2]	Focus on Product Reviews. Builds a Lexicon to predict polarity of product features phrases that are summarized to provide an overall score to that product feature.	✓	
Opinion Finder (MPQA) [22, 23]	Performs subjectivity analysis trough a framework with lexical analysis former and a machine learning approach latter.	✓	✓
SentiWordNet [24, 25]	Construction of a lexical resource for Opinion Mining based on WordNet [26]. The authors grouped adjectives, nouns, etc. in synonym sets (synsets) and associated three polarity scores (positive, negative and neutral) for each one.	✓	✓
LIWC [7]	An acronym for Linguistic Inquiry and Word Count, LIWC is a text analysis paid tool to evaluate emotional, cognitive, and structural components of a given text. It uses a dictionary with words classified into categories (anxiety, health, leisure, etc.). An updated version was launched in 2015.	✓	
Sentiment140 [27]	Sentiment140 (previously known as 'Twitter Sentiment') was proposed as an ensemble of three classifiers (Naive Bayes, Maximum Entropy, and SVM) built with a huge amount of tweets containing emoticons collected by the authors. It has been improved and transformed into a		
SenticNet [28]		✓	
AFINN [29] - a new ANEW	Builds a Twitter-based sentiment Lexicon including Internet slangs and obscene words. AFINN can be considered as an expansion of ANEW [30], a dictionary created to provides emotional ratings for English words. ANEW dictionary rates words in terms of pleasure, arousal and dominance.	✓	

So many options!

SO-CAL [31]	Creates a new Lexicon with unigrams (verbs, adverbs, nouns and adjectives) and multi-grams (phrasal verbs and intensifiers) hand ranked with scale +5 (strongly positive) to −5 (strongly negative). Authors also included part of speech processing, negation and intensifiers.	✓	
Emoticons DS (Distant Supervision) [32]	Creates a scored lexicon based on a large dataset of tweets. Its based on the frequency each lexicon occurs with positive or negative emotions.	✓	
NRC Hashtag [33]	Builds a lexicon dictionary using a Distant Supervised Approach. In a nutshell it uses known hashtags (i.e. #joy, #happy, etc.) to classify the tweet. Afterwards, it verifies frequency each specific n-gram occurs in a emotion and calculates its Strong of Association with that emotion.	✓	
Pattern.en [34]	Python Programming Package (toolkit) to deal with NLP, Web Mining and Sentiment Analysis. Sentiment analysis is provided through averaging scores from adjectives in the sentence according to a bundle lexicon of adjective.	✓	
SASA [35]	Detects public sentiments on Twitter during the 2012 U.S. presidential election. It is based on the statistical model obtained from the classifier Naive Bayes on unigram features. It also explores emoticons and exclamations.		✓
PANAS-t [8]	Detects mood fluctuations of users on Twitter. The method consists of an adapted version (PANAS) Positive Affect Negative Affect Scale [36], well-known method in psychology with a large set of words, each of them associated with one from eleven moods such as surprise, fear, guilt, etc.	✓	
Emolex [37]	Builds a general sentiment Lexicon crowdsourcing supported. Each entry lists the association of a token with 8 basic sentiments: joy, sadness, anger, etc. defined by [38]. Proposed Lexicon includes unigrams and bigrams from Macquarie Thesaurus and also words from GI and WordNet.	✓	
USent [39]	Infer additional reviews user ratings by performing sentiment analysis (SA) of user comments and integrating its output in a nearest neighbor (NN) model that provides multimedia recommendations over TED talks.	✓	✓

Data Analysis: Sentiment Analysis

- Can be as simple as counting the positive and negative words and normalize by number of words in a text
- usually computed on a range:
  - -1 (negative) to 1 (positive)
  - very negative, negative, neutral / none, positive, positive
  - VADER: positive score, negative score, neutral score and *complex* score where one can fix a threshold
  - usually: negative < -0.1, 0.1 < positive

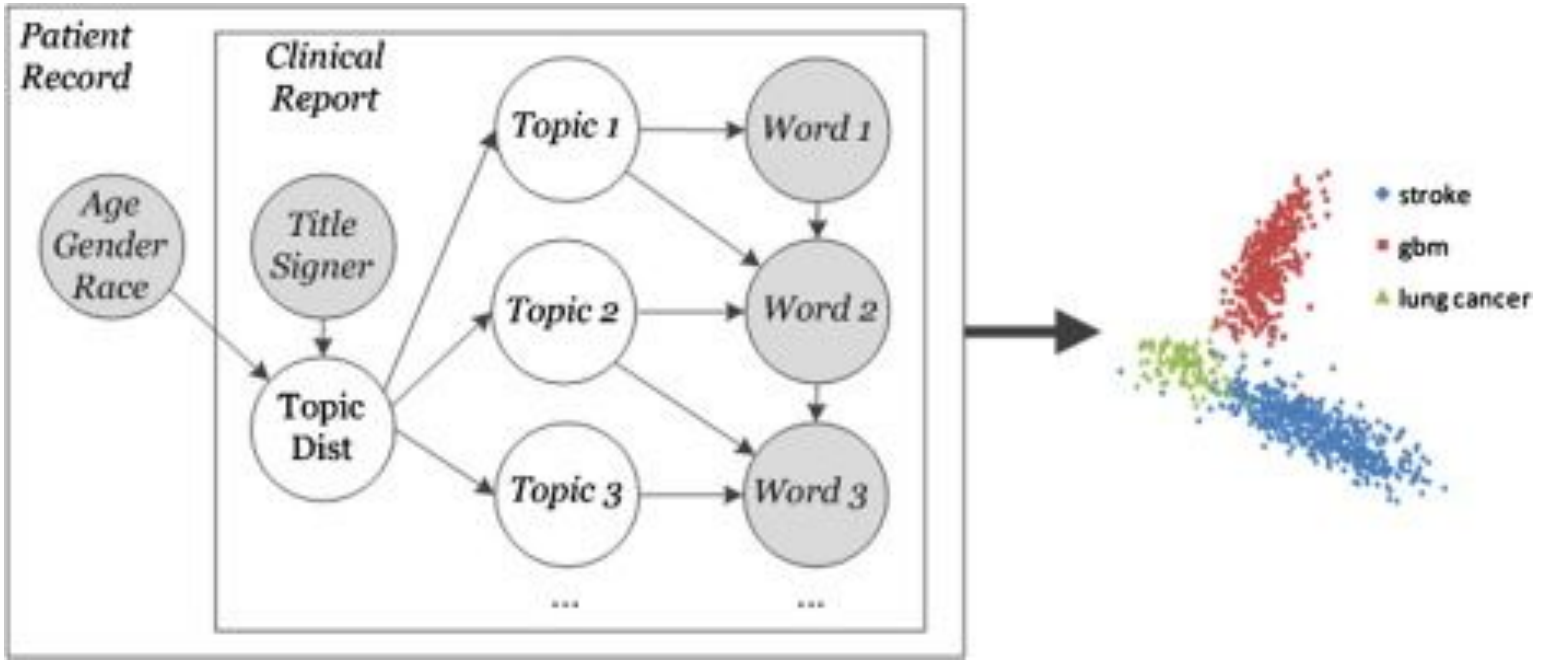
Data Analysis: Sentiment Analysis

- VADER: Valence Aware Dictionary for sEntiment Reasoning
- Gold-standard sentiment dictionary + preprocessing engine

VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

C.J. Hutto **Eric Gilbert**
Georgia Institute of Technology, Atlanta, GA 30032
cjhutto@gatech.edu gilbert@cc.gatech.edu

Data Analysis: Topic Modeling



Data Analysis: Topic Modeling

Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data

PABLO BARBERÁ *University of Southern California*
ANDREU CASAS *New York University*
JONATHAN NAGLER *New York University*
PATRICK J. EGAN *New York University*
RICHARD BONNEAU *New York University*
JOHN T. JOST *New York University*
JOSHUA A. TUCKER *New York University*

Are legislators responsive to the priorities of the public? Research demonstrates a strong correspondence between the issues about which the public cares and the issues addressed by politicians, but conclusive evidence about who leads whom in setting the political agenda has yet to be uncovered. We answer this question with fine-grained temporal analyses of Twitter messages by legislators and the public during the 113th U.S. Congress. After employing an unsupervised method that classifies tweets sent by legislators and citizens into topics, we use VAR models to explore whose priorities more strongly predict the relationship between citizens and politicians. We find that legislators are more likely to follow, than to lead, discussion of public issues, results that hold even after controlling for the agenda-setting effects of the media. We also find, however, that legislators are more

Data Analysis: Topic Modeling

Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data

PABLO B
ANDREU
JONATH
PATRICK
RICHAR
JOHN T.
JOSHUA

A re
st
is
setting the
temporal
Congress.
and citizen
the relatio
to follow,
for the age

Modeling of Political Discourse Framing on Twitter

Kristen Johnson, Di Jin, Dan Goldwasser
Department of Computer Science
Purdue University, West Lafayette, IN 47907
{john1187, jind, dgoldwas}@purdue.edu

Abstract

Framing is a political strategy in which politicians carefully word their statements in order to control public perception and discussion of current issues. Previous works exploring political framing have focused on analysis of frames in longer texts, such as newspaper articles, or tweets relevant to specific events. We present the first in-depth analysis of issue-independent framing for political discourse in social media, specifically the microblogging platform Twitter. Building upon the fifteen frames designed by Boydstun, we propose three additional frames relevant to Twitter and provide insights into the dynamic usage of frames by party and over time. Finally we present a global probabilistic model for combining linguistic, issue, and party bias features of the tweets of politicians for the task of tweet frame prediction.

et al.), tweet policy issues, user party affiliation, and frequent phrases used by politicians on Twitter. These indicators are extracted via weakly supervised models and then declaratively combined into a global model using Probabilistic Soft Logic (PSL), a recently introduced probabilistic modeling framework (Bach et al. 2013). PSL specifies high level rules over a relational representation of these features, which are compiled into a graphical model called a hinge-loss Markov random field that is used to make the frame prediction.

In summary, this paper makes the following contributions: (1) This work is among the first to look into general framing analysis of U.S. politicians on Twitter. Extending the annotation guidelines of Boydstun et al., we annotated 2,050 tweets, a subset of our total evaluation set of 92,457 tweets,

Data Analysis: Topic Modeling

All topic models are based on the same basic assumption:

- each **document** consists of a mixture of ***topics***, and
- each ***topic*** consists of a collection of **words**

Data Analysis: Topic Modeling

- The semantics of our document are actually being governed by some **hidden**, or "**latent**" variables that we are not observing

- Goal of topic modeling: uncover these latent variables — *topics*

Data Analysis: Toxicity

Data Analysis: Toxicity

Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Abstract

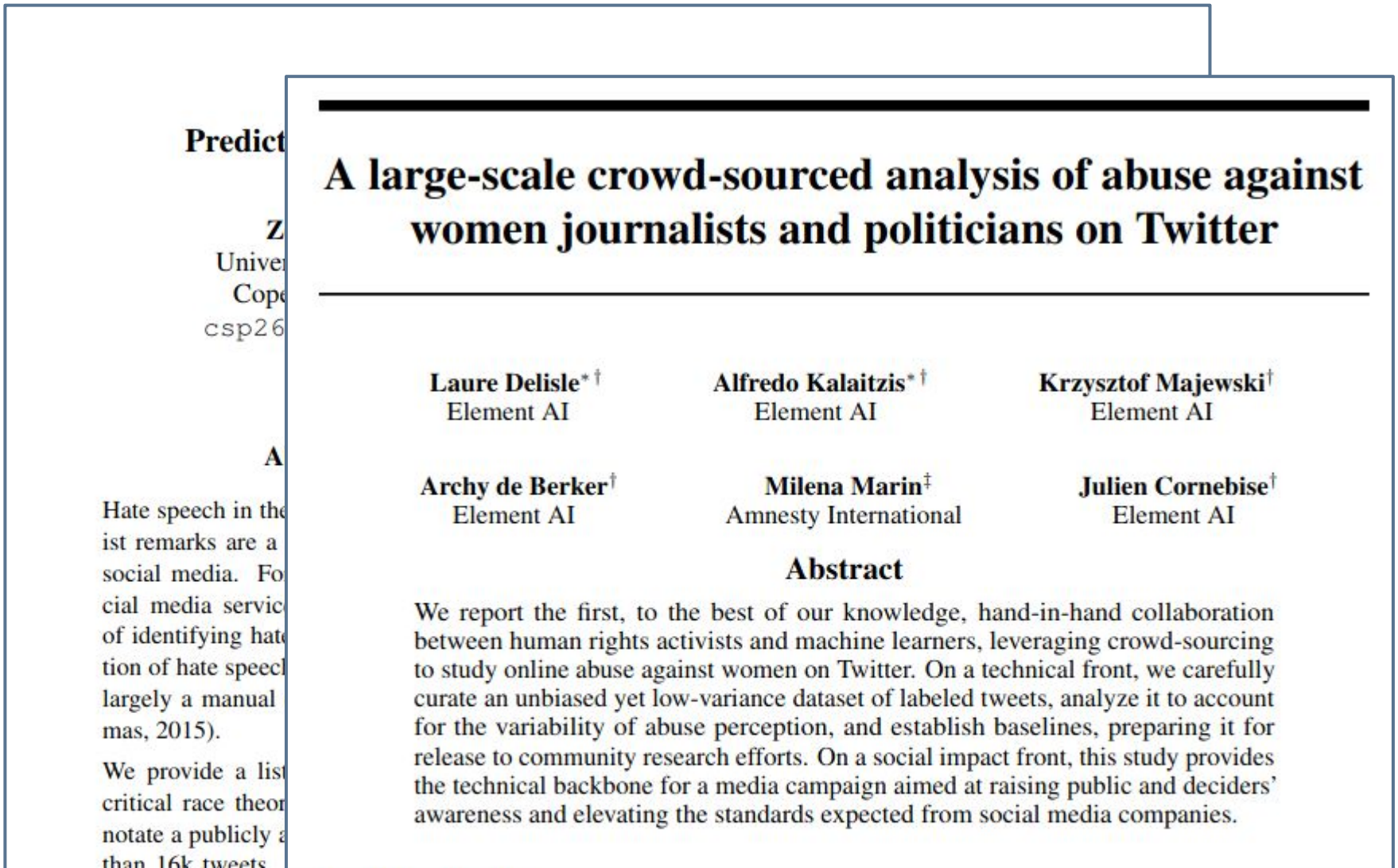
Hate speech in the form of racist and sexist remarks are a common occurrence on social media. For that reason, many social media services address the problem of identifying hate speech, but the definition of hate speech varies markedly and is largely a manual effort (BBC, 2015; Lomas, 2015).

We provide a list of criteria founded in critical race theory, and use them to annotate a publicly available corpus of more

much of this moderation requires manual review of questionable documents, which not only limits how much a human annotator can be reviewed, but also introduces subjective notions of what constitutes hate speech. A reaction to the "Black Lives Matter" movement, a campaign to highlight the devaluation of lives of African-American citizens sparked by extrajudicial killings of black men and women (Matter, 2012), at the Facebook campus shows how individual biases manifest in evaluating hate speech (Wong, 2016).

In spite of these reasons, NLP research on hate speech has been very limited, primarily due to the

Data Analysis: Toxicity



Predict

Z
Univer
Cope
csp26

A

Hate speech in the
ist remarks are a
social media. Fo
cial media servic
of identifying hat
tion of hate speec
largely a manual
mas, 2015).

We provide a list
critical race theor
notate a publicly a
than 16k tweets

A large-scale crowd-sourced analysis of abuse against women journalists and politicians on Twitter

Laure Delisle[*†] Element AI
Alfredo Kalaitzis[*†] Element AI
Krzysztof Majewski[†] Element AI
Archy de Berker[†] Element AI
Milena Marin[‡] Amnesty International
Julien Cornebise[†] Element AI

Abstract

We report the first, to the best of our knowledge, hand-in-hand collaboration between human rights activists and machine learners, leveraging crowd-sourcing to study online abuse against women on Twitter. On a technical front, we carefully curate an unbiased yet low-variance dataset of labeled tweets, analyze it to account for the variability of abuse perception, and establish baselines, preparing it for release to community research efforts. On a social impact front, this study provides the technical backbone for a media campaign aimed at raising public and deciders' awareness and elevating the standards expected from social media companies.

Data Analysis: Toxicity

- From Google's Perspective API: https://github.com/conversationai/perspectiveapi
- "Perspective is an API that uses machine learning models to score the perceived impact a comment might have on a conversation."
- We will be looking at one model in particular: 'toxicity'
- Toxic is defined as... "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Data Analysis: Toxicity

The Risk of Racial Bias in Hate Speech Detection

Maarten Sap[◇] **Dallas Card**[♣] **Saadia Gabriel**[◇] **Yejin Choi**[◇♡] **Noah A. Smith**[◇♡]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
ment, Carnegie Mellon University, Pittsburgh, USA
for Artificial Intelligence, Seattle, USA
p@cs.washington.edu

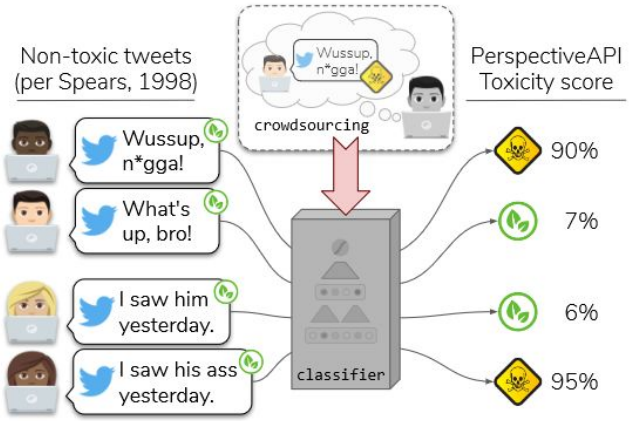
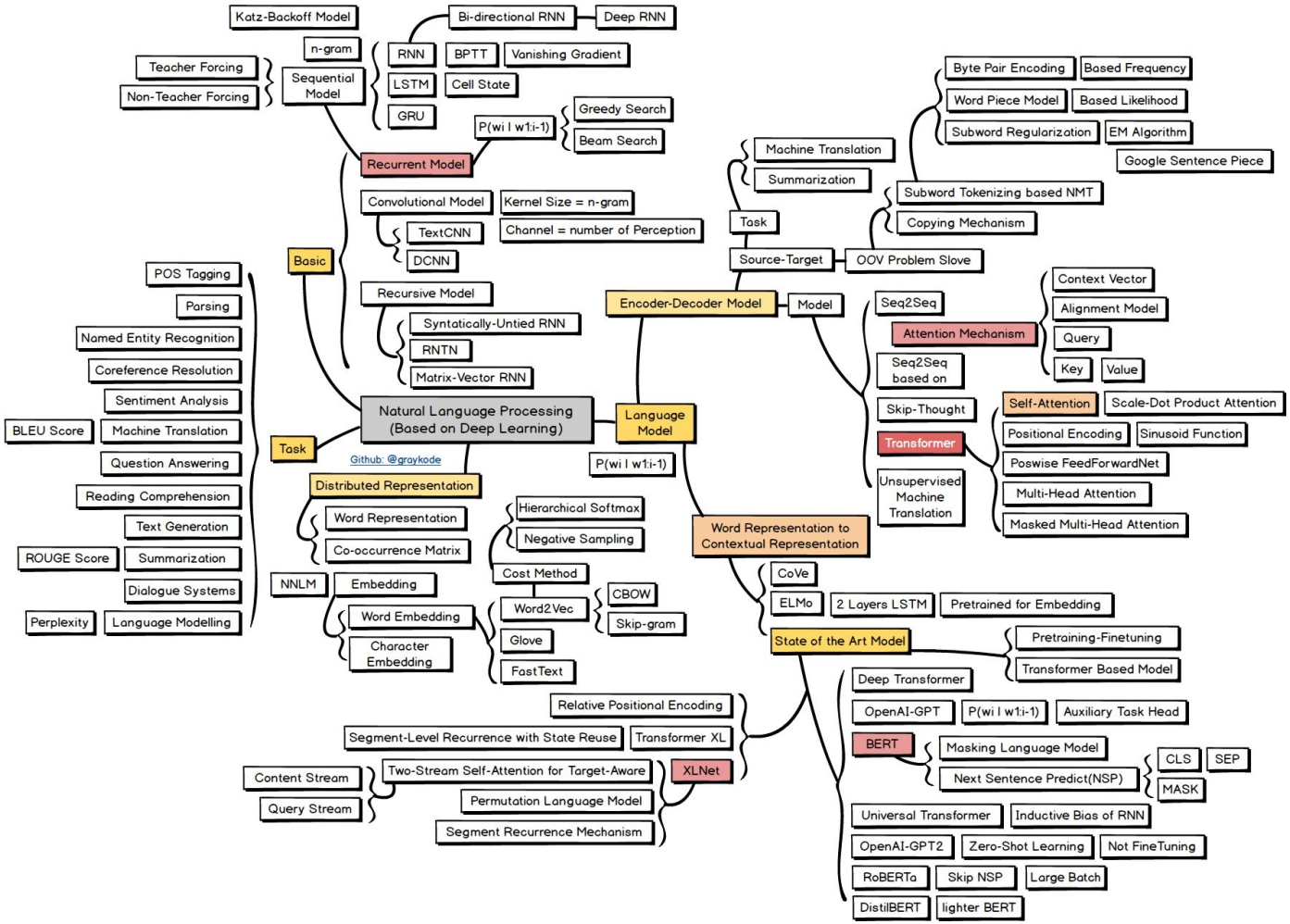


Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

Data Analysis: Other Ideas

- Named Entity Recognition
- Social Category analysis:
  a. demographics
  b. political ideology
  c. mental health
- Time Series analysis: How does a property change over time?
- Engagement analysis: What kind of content gets the most favorites, like, retweets, upvotes, score, etc?
- Combinations: Which topics have most negative sentiment?

Data Analysis: Other Ideas



Data Analysis: Other Ideas

- Word Embeddings:
- BERT: Deep Learning method which can be used for multiple Linguistic tasks: https://arxiv.org/abs/1810.04805
- Beyond English: https://github.com/google-research/bert/blob/master/multilingual.md
- General NLP overview:
  a. https://nlpprogress.com/
  b. https://github.com/graykode/nlp-roadmap
  c. https://lena-voita.github.io/nlp_course.html

Data Analysis: Other Ideas

- Beyond Text:
  a. Images
     i. Object Recognition
     ii. Scene description / captioning
  b. Network
     i. Who follows whom?
     ii. Community analysis
     iii. Hubs and authorities