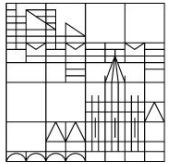
# Text-as-Data Methods

Indira Sen

03.06.24

Universität
Konstanz



# Agenda

- Recap: Exploratory Data Analysis and Data Visualization
- Today: Text-as-data methods
  - Supervised techniques
    - Sentiment Analysis Applications
    - Training your own models
    - Natural language inference for text classification
    - Validation
  - Unsupervised techniques
    - Topic Modeling Applications
    - Validation

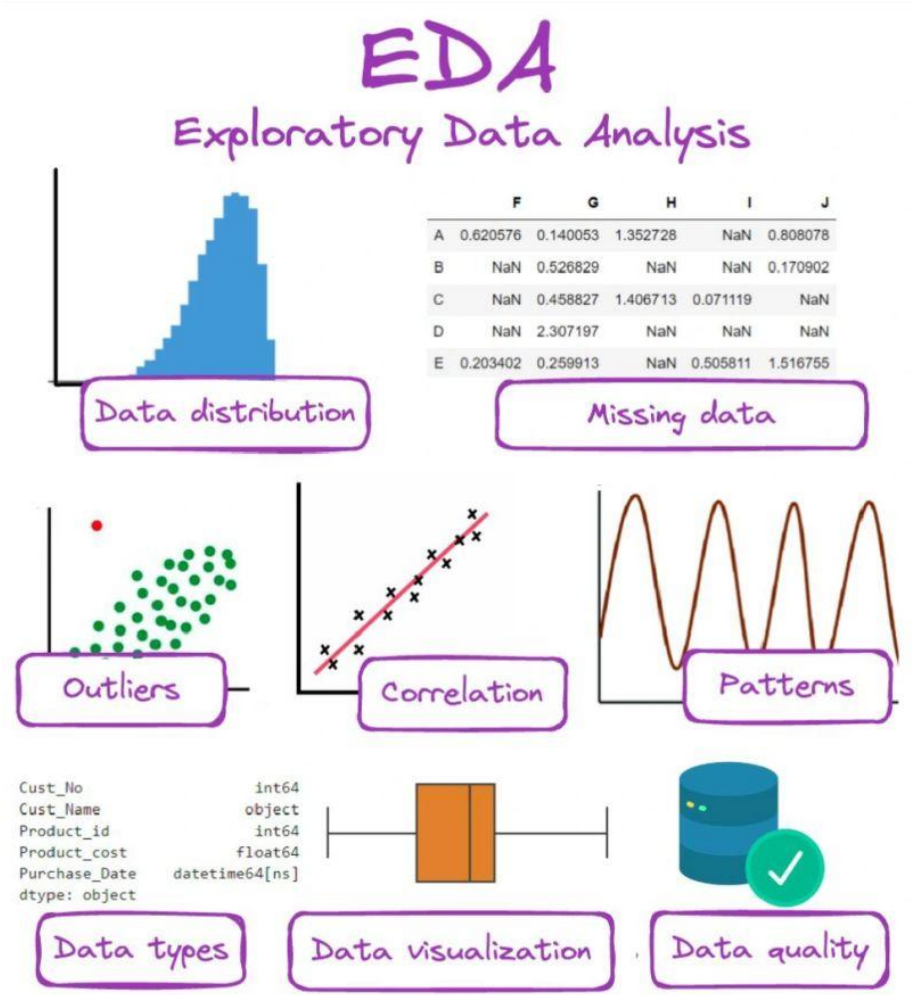
# Recap

# Data Cleaning

1. Remove unnecessary items from our dataset
    a. function words (‘the’, ‘on’, etc)
2. Maintain order and consistency.
3. Standardization, e.g., time formats
4. Deduplication: not just exact match but also ‘near-duplicates’

# 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

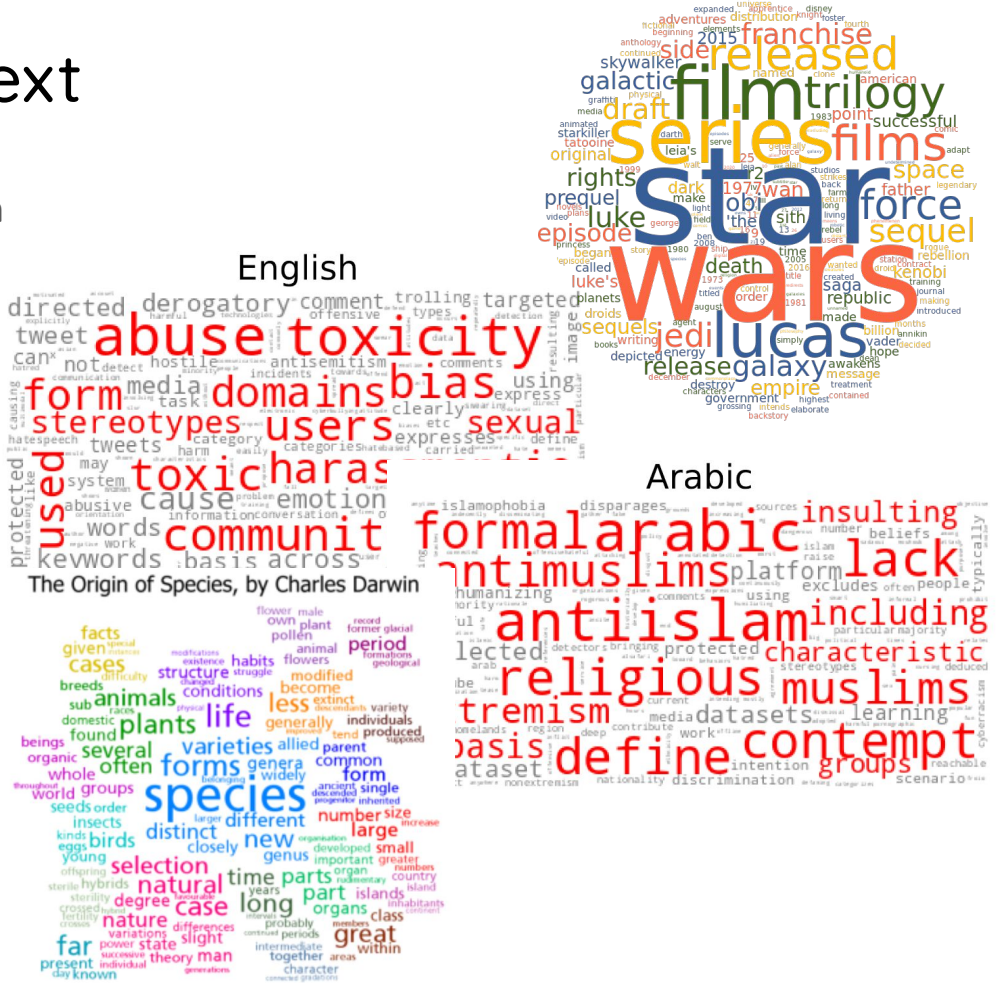



https://www.markovml.com/blog/exploratory-data-analysis

# Data Visualization for Text

- Word clouds: easy to make, but often too vague.

- Variants of word clouds
  - Accentuated word clouds
  - Semantically grouped word clouds



# Today: Text-as-Data

# Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]

# Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
  - What are topics that lead to arguments in long-term relationships?
  - How do citizens perceive the performance of politicians during the pandemic?

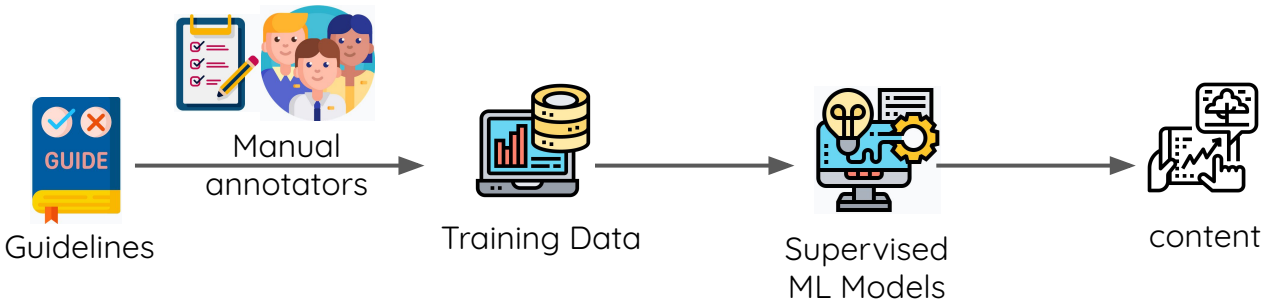
# Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
    NYtimes articles frames [Mendelsohn’21]
  - What are topics that lead to arguments in long-term relationships?
    r/relationshipadvice posts topics
  - How do citizens perceive the performance of politicians during the pandemic?
    Twitter posts stance

# Content Analysis Pipeline using ML/NLP

- unsupervised

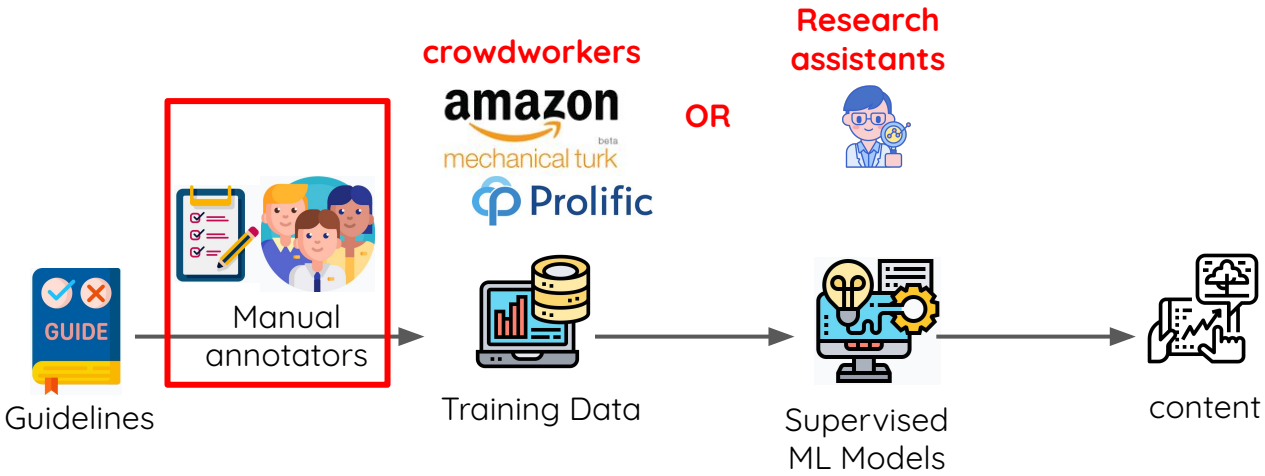

- supervised



# Content Analysis Pipeline using ML/NLP

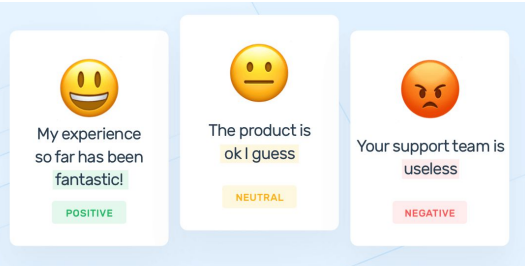
- unsupervised

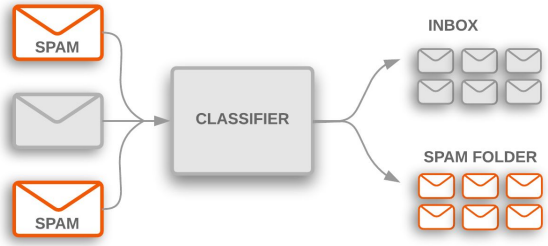

- supervised



# Text Classification

- One of the most basic and popular NLP tasks
- Many connections to content analysis
- Supervised Machine Learning





Check out https://lena-voita.github.io/nlp_course/text_classification.html for a deeper dive into text classification

# Sentiment 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

# Sentiment Analysis

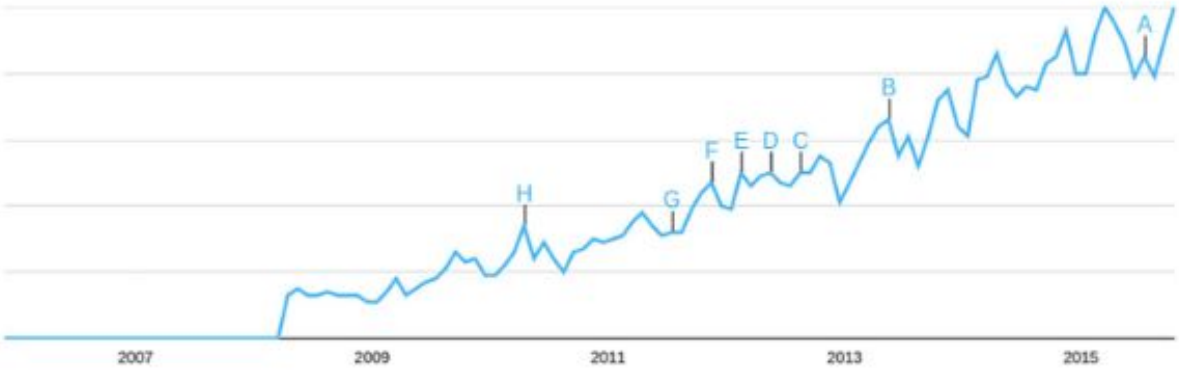



**Figure 1 Searches on Google for the Query: 'Sentiment Analysis'.** This figure shows the steady growth on the number of searches on the topic, according to Google Trends, mainly after the popularization of online social networks (OSNs).

# Sentiment Analysis

## From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

**Brendan O'Connor**[†] **Ramnath Balasubramanyan**[†] **Bryan R. Routledge**[§] **Noah A. Smith**[†]
brenocon@cs.cmu.edu rbalasub@cs.cmu.edu routledge@cmu.edu nasmith@cs.cmu.edu

[†]School of Computer Science
Carnegie Mellon University

[§]Tepper School of Business

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.



**Abstract**

We connect measures of public opinion measured from polls with sentiment measured from text. We analyze several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and find they correlate to sentiment word frequencies in contemporaneous Twitter messages. While our results vary across

statistics derived from extremely simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion, and can also predict future movements in the polls. We find that temporal smoothing is a critically important issue to support a successful model.

# Sentiment Analysis

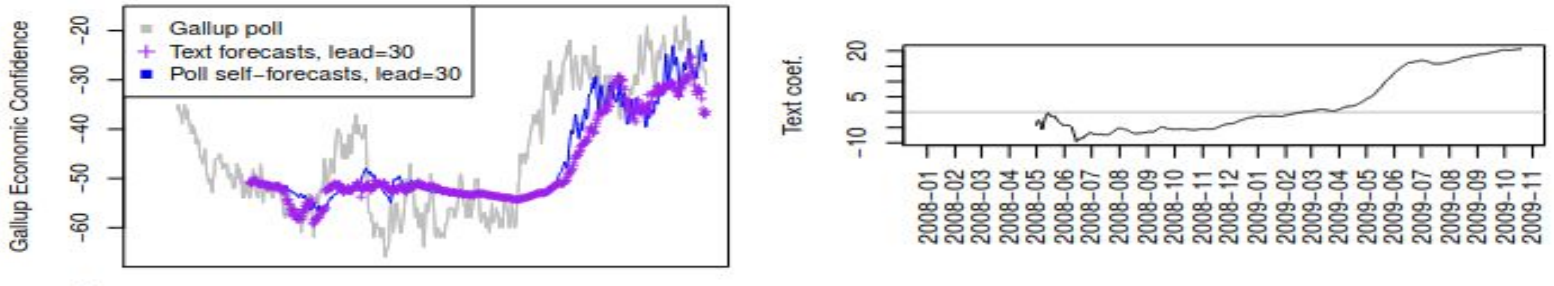



Figure 8: Rolling text-based forecasts (above), and the text sentiment ($MA_t$) coefficients $a$ for each of the text forecasting models over time (below).

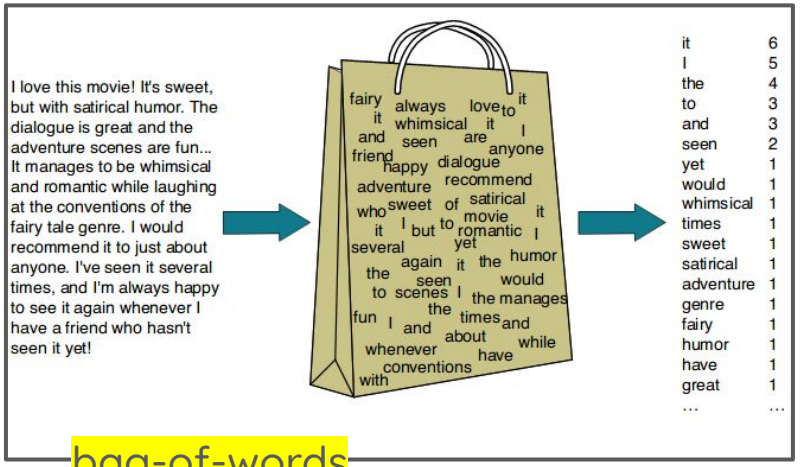
# Sentiment Analysis

- 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

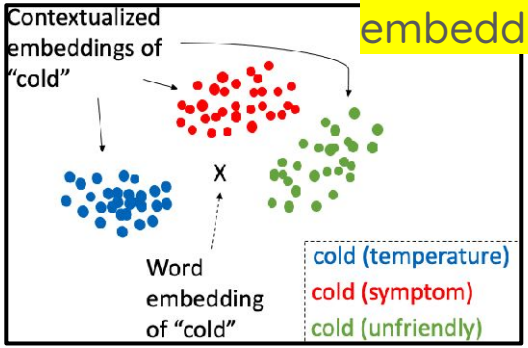
# Text **Classification**

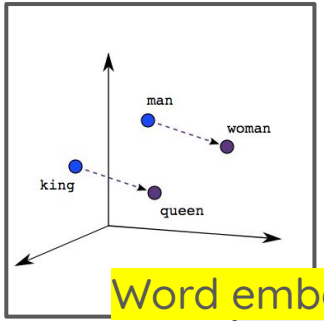
- Given ‘something’ (a document, a number, a set of numbers, etc), classify it based on a **fixed** number of categories (‘classes’)
- Numerical things are easy to classify because computers know bits (0s and 1s)
- Therefore, we turn words to numbers
  - Obtain a **representation**
  - Classify

# Text **Classification**

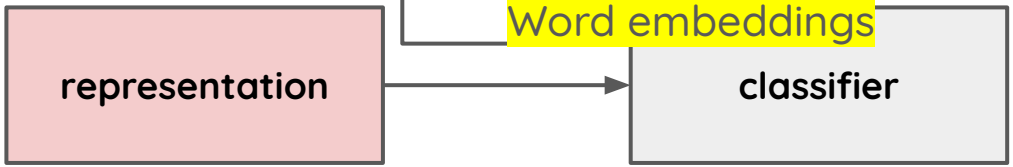



Contextualized word embeddings

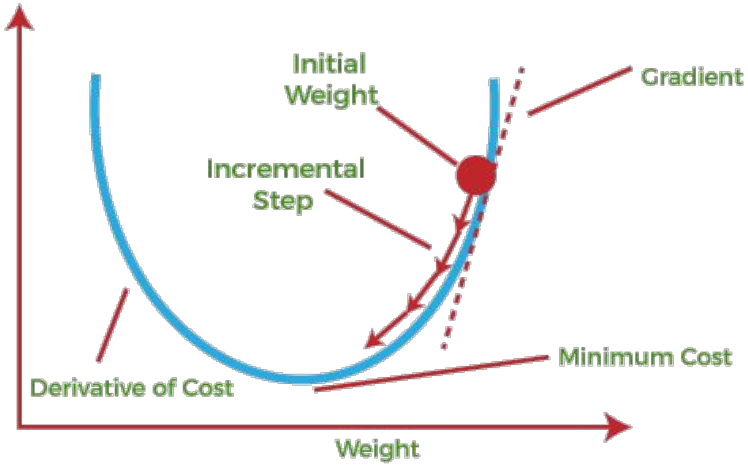


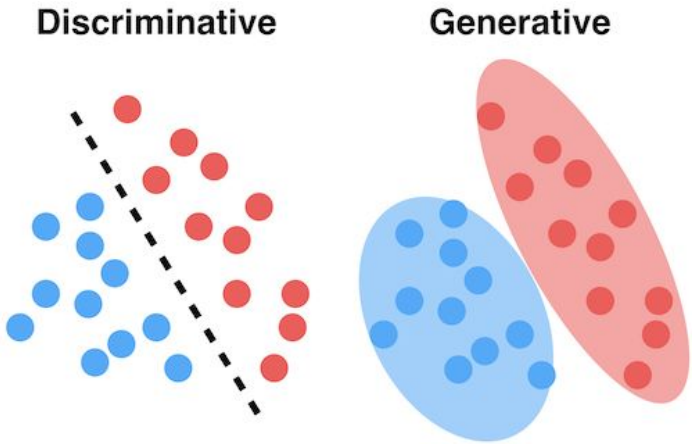




Word embeddings



# Text **Classification**



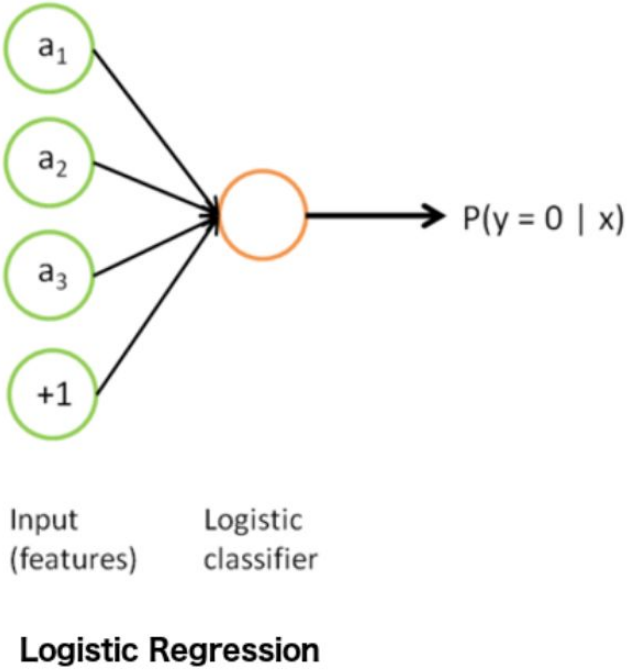





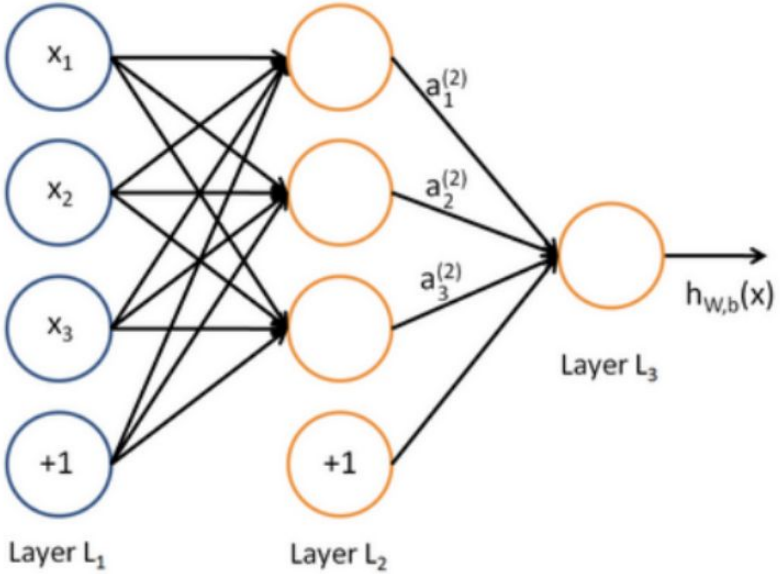
Find the **parameters** of the representation which minimizes error

# Neural Networks and Deep Learning

No need for feature engineering, the network learns higher or lower dimensions on itself




**Logistic Regression**




**Neural Network**

# Neural Networks and Deep Learning

What do we mean by feature engineering?

Say for sentiment analysis,

Table II. Summary of the Articles Employed a Supervised Method to Address TSA

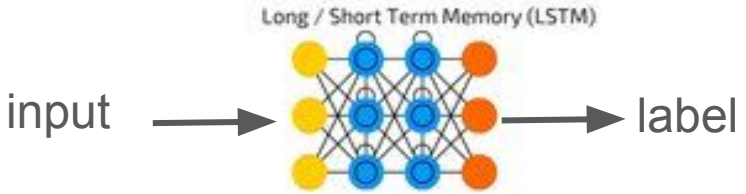
| Study | Task | Algorithms | Features | Dataset |
|---|---|---|---|---|
| Go et al. [2009] | TSA | NB, MaxEnt, SVM | unigrams, bigrams, POS | STS |
| Pak and Paroubek [2010] | TSA | MNB, SVM, CRF | unigrams, bigrams, trigrams, POS | own |
| Barbosa and Feng [2010] | TSA | SVM | meta-features (POS, polarity-MPQA), tweet syntax (i.e., retweet, hashtags, emoticons, links etc.) | own |
| Davidov et al. [2010] | TSA | kNN | word and n-gram based, punctuation-based, pattern-based | OC |
| Bakliwal et al. [2012] | TSA | SVM, NB | words' polarity, unigrams, bigrams, emoticons, hashtags, URLs, targets etc. | STS, Mejaj [Bora 2012] |
| Mohammad et al. [2013] | TSA | SVM | word/character n-grams, POS, caps, lexicons, punctuation, negation, tweet-based | SemEval-2013 |
| Kiritchenko et al. [2014] | TSA | linear kernel SVM, MaxEnt | word/character n-grams, POS, caps, punctuation, emoticons, automatic sentiment lexicons, polarity, emphatic lengthening | SemEval-2013 |
| Asiaee et al. [2012] | TSA | dictionary learning, WSVM, NB, kNN, | | DETC |
| Agarwal et al. [2011] | TSA | SVM | POS, unigrams, DAL lexicon, caps, exclamation etc. | own |
| Aisopos et al. [2011] | TSA | MNB, C4.5 tree | n-grams | own |
| Kouloumpis et al. [2011] | TSA | AdaBoost | N-gram with lexicon features, twitter-based, POS | STS, ETC |
| Saif et al. [2012b] | TSA | NB | unigrams, POS, sentiment-topic, semantic features | STS, HCR, OMD |
| Hamdan et al. [2013] | TSA | SVM, NB | unigrams, concepts (DBPedia), verb groups/adjectives (WordNet) and senti-features (SentiWordNet) | SemEval-2013 |
| Jiang et al. [2011] | entity-TSA | SVM | unigrams, emoticons, hashtags, punctuation the General Inquirer lexicon | own |
| Aston et al. [2014] | TSA | Perceptron with Best Learning | character n-grams | Sanders |

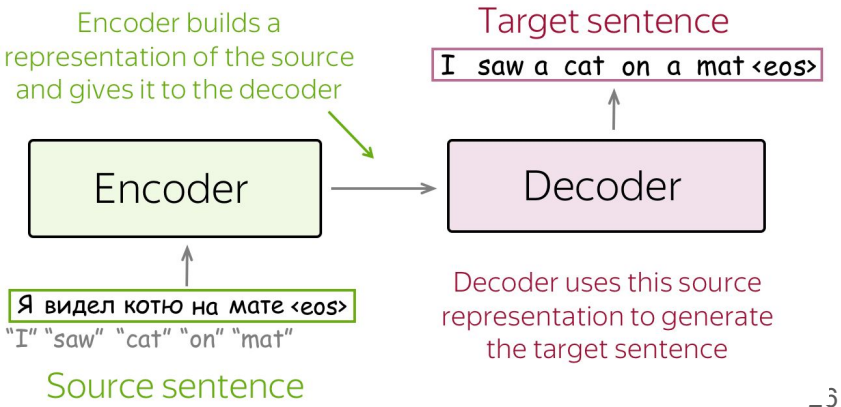
Like It or Not: A Survey of Twitter Sentiment Analysis Methods, Anastasia Giachanou And Fabio Crestani, 2016

Method

# Deep Learning in NLP (till mid 2010's)

- Precludes need for explicit feature engineering
- Automatically creates complex representations
- Popular Classifiers are RNNs (GRUs and LSTMs)
- Classification is one strand of problems
- Another: sequence-to-sequence output, e.g., Machine Translation




- Lena Voita, NLP Course for you

# Transformers

- RNNs in architectures have limited context
- Fix: replace all components with attention

| | Seq2seq without attention | Seq2seq with attention | Transformer |
|---|---|---|---|
| processing within encoder | RNN/CNN | RNN/CNN | attention |
| processing within decoder | RNN/CNN | RNN/CNN | attention |
| decoder-encoder interaction | static fixed-sized vector | attention | attention |



- Lena Voita, NLP Course for you

# BERT, GPT (the first one), ….

What is encoded

How it is used for downstream tasks

**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

**Jacob Devlin** **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

**Abstract**

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the

# The old vs. ‘new’ paradigm

**OLD:** Train a model on a task where we have labeled data

**NEW:**

1. **Pre-train** a model on a task where we have lots of data
   - *e.g., we have lots and lots of internet data*


2. **Fine-tune** the model on your downstream task
   - *e.g., a specific, curated dataset that you’re studying*

# What is fine-tuning?

- Inputs:
  - pretrained model (usually a transformer, BERT and variants are still very popular)
  - Labeled dataset (rule-of-thumb, at least 1K examples)
- Adds a new layer containing classification parameters.
- All model parameters are updated to maximize the log probability of the labels.
- Output: classifier for your specific task

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised


Guidelines

Manual annotators

Training Data

Supervised ML Models

content

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised


Guidelines

Manual annotators

Training Data

Supervised ML Models

content

- ‘off-the-shelf’

Off-the-shelf classifier

content

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised


Guidelines

Manual annotators

Training Data

Supervised ML Models

content

- ‘off-the-shelf’

**Created for other contexts**

**Still need to validate if it works well**

Off-the-shelf classifier

content

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised


Guidelines

Manual annotators

Training Data

Supervised ML Models

content

- ‘off-the-shelf’ e.g., MNLI, VADER for sentiment

**Created for other contexts**

**Still need to validate if it works well**

Off-the-shelf classifier

content

# Zero-Shot Natural Language Inference (NLI)

- The NLP task of detecting if a given premise justifies the given hypothesis
- Also called textual entailment
- Typically there are three classes:
  - Entailment
  - Contradiction
  - Neutral or none

| Premise | Relation | Hypothesis |
|---|---|---|
| A turtle danced. | entails | A turtle moved. |
| turtle | contradicts | linguist |
| Every reptile danced. | neutral | A turtle ate. |
| Some turtles walk. | contradicts | No turtles move. |
| James Byron Dean refused to move without blue jeans. | entails | James Dean didn't dance without pants. |
| Mitsubishi Motors Corp's new vehicle sales in the US fell 46 percent in June. | contradicts | Mitsubishi's sales rose 46 percent. |
| Acme Corporation reported that its CEO resigned. | entails | Acme's CEO resigned. |

https://web.stanford.edu/class/cs224u/2021/slides/cs224u-2021-nli-part1-handout.pdf

# Zero-Shot Natural Language Inference (NLI)

- NLI is helpful for many other tasks:
  - Question answering
  - Summarization
  - …
- And now more recently, for zero-shot text classification
- Premise: Joe Biden's inauguration…
- Hypothesis: This example is about **politics**.

**Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach**

**Wenpeng Yin, Jamaal Hay, Dan Roth**
Cognitive Computation Group
Department of Computer and Information Science, University of Pennsylvania
{wenpeng,jamaalh,danroth}@seas.upenn.edu

**Abstract**

Zero-shot text classification (0SHOT-TC) is a challenging NLU problem to which little attention has been paid by the research community. 0SHOT-TC aims to associate an appropriate label with a piece of text, irrespective of the text domain and the aspect (e.g., topic, emotion, event, etc.) described by the label. And there are only a few articles studying 0SHOT-TC, all focusing only on topical categorization which, we argue, is just the tip of the iceberg in 0SHOT-TC. In addition, the chaotic experiments in literature make no uniform comparison, which blurs the progress.

This work benchmarks the 0SHOT-TC problem

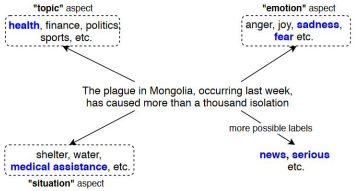



Figure 1: A piece of text can be assigned labels which describe the different aspects of the text. Positive labels are in blue.

has attracted little attention despite its great po-

[Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach]

# 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

# 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 n... ANEW | ...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

- VADER: Valence Aware Dictionary for sEntiment Reasoning
- Gold-standard sentiment dictionary + preprocessing engine
- positive score, negative score, neutral score and *complex* score where one can fix a threshold
- usually: negative < -0.1, 0.1 < positive

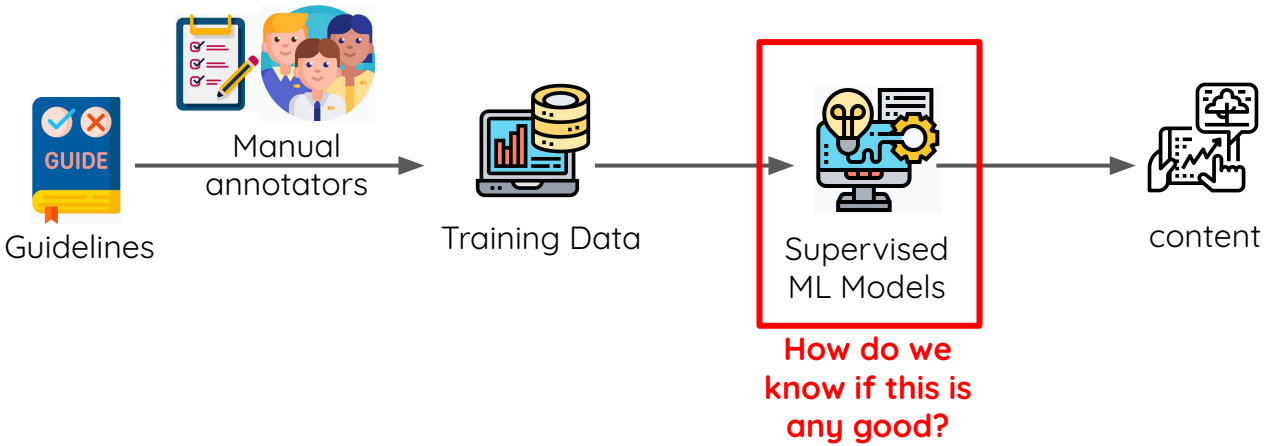
**VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text**

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

# Validating supervised classification

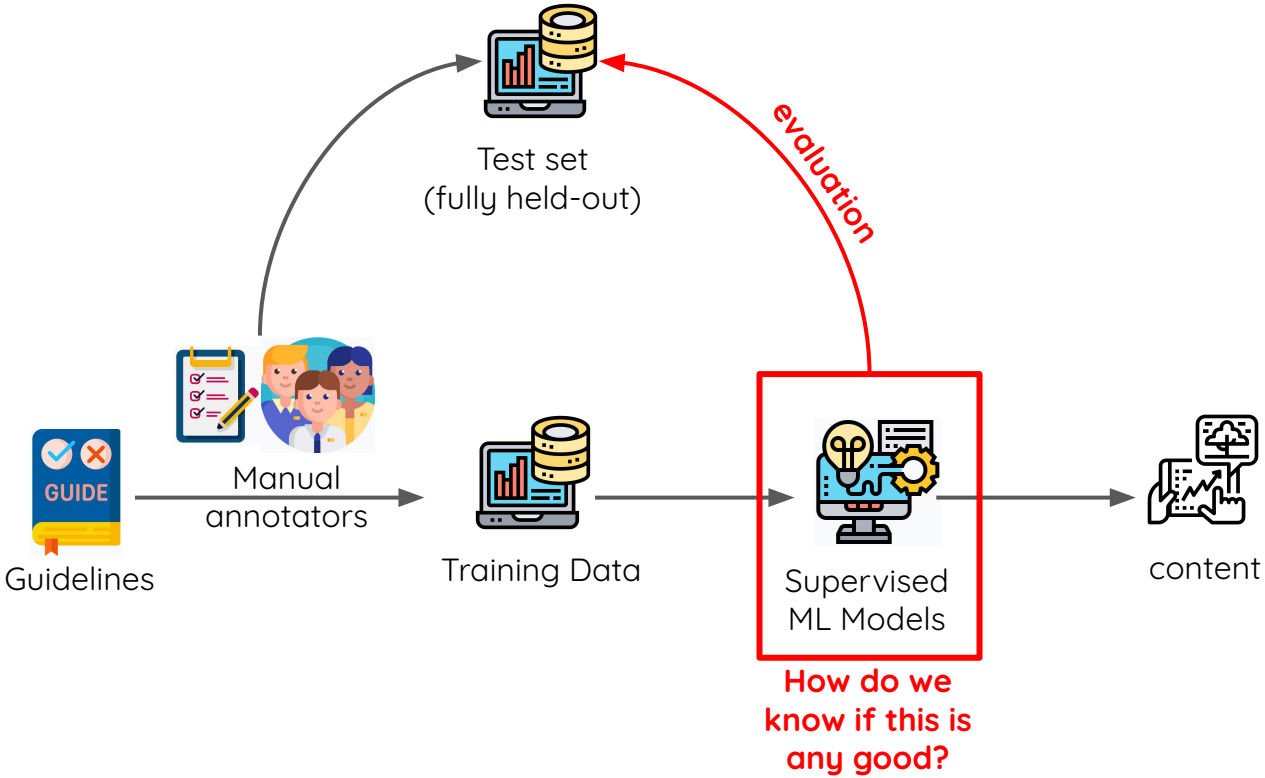
# Modern Text Classification Pipeline

- supervised



# Modern Text Classification Pipeline

- supervised



# Evaluating Text Classification

| | | Predicted condition | | | |
|---|---|---|---|---|---|
| | Total population = P + N | **Positive (PP)** | **Negative (PN)** | Informedness, bookmaker informedness (BM) = TPR + TNR − 1 | Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$ |
| Actual condition | **Positive (P)** | **True positive (TP),** hit | **False negative (FN),** type II error, miss, underestimation | True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$ | False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$ |
| | **Negative (N)** | **False positive (FP),** type I error, false alarm, overestimation | **True negative (TN),** correct rejection | False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$ | True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$ |
| | Prevalence $= \frac{P}{P + N}$ | Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$ | False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$ | Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$ | Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$ |
| | Accuracy (ACC) $= \frac{TP + TN}{P + N}$ | False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$ | Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$ | Markedness (MK), deltaP (Δp) = PPV + NPV − 1 | Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$ |
| | Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$ | $F_1$ score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$ | Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$ | Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$ | Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$ |

Sources: [4][5][6][7][8][9][10][11][12] view · talk · edit

https://en.wikipedia.org/wiki/Precision_and_recall

# Evaluating Text Classification

| | | Predicted condition | | | |
|---|---|---|---|---|---|
| | Total population = P + N | **Positive (PP)** | **Negative (PN)** | Informedness, bookmaker informedness (BM) = TPR + TNR − 1 | Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$ |
| Actual condition | **Positive (P)** | **True positive (TP),** hit | **False negative (FN),** type II error, miss, underestimation | True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$ | False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$ |
| | **Negative (N)** | **False positive (FP),** type I error, false alarm, overestimation | **True negative (TN),** correct rejection | False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$ | True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$ |
| | Prevalence $= \frac{P}{P+N}$ | Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$ | False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$ | Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$ | Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$ |
| | Accuracy (ACC) $= \frac{TP + TN}{P + N}$ | False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$ | Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$ | Markedness (MK), deltaP (Δp) = PPV + NPV − 1 | Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$ |
| | Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$ | $F_1$ score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$ | Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$ | Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$ | Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$ |

Sources: [4][5][6][7][8][9][10][11][12] view · talk · edit

https://en.wikipedia.org/wiki/Precision_and_recall

# Unsupervised Methods: Topic Modeling

## Topic Modeling

Topic modeling is **unsupervised**, meaning it **doesn't require labeled training data.** Machine Learning algorithms simply look at the documents and group them into topics.

Topic modeling can find **hidden structures** in text data.

Topic modeling is more flexible than text classification since it **doesn't require a predefined set of categories**. Which can be time-consuming and expensive without adequate tools.

## Text Classification

Text classification uses **supervised** Machine Learning techniques and **labeled training data** to learn how to classify text.

Text classification is more **focused on assigning labels to texts.**

Text classification **requires pre-labeling** of the data,

https://levity.ai/blog/what-is-topic-modeling

# Topic Modeling for Social Research

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

Barberá, Pablo, et al. "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data." *American Political Science Review* 113.4 (2019): 883-901.

JONATH
PATRICK
RICHAR
JOHN T.
JOSHUA

A re
st
is
setting the

## Modeling of Political Discourse Framing on Twitter

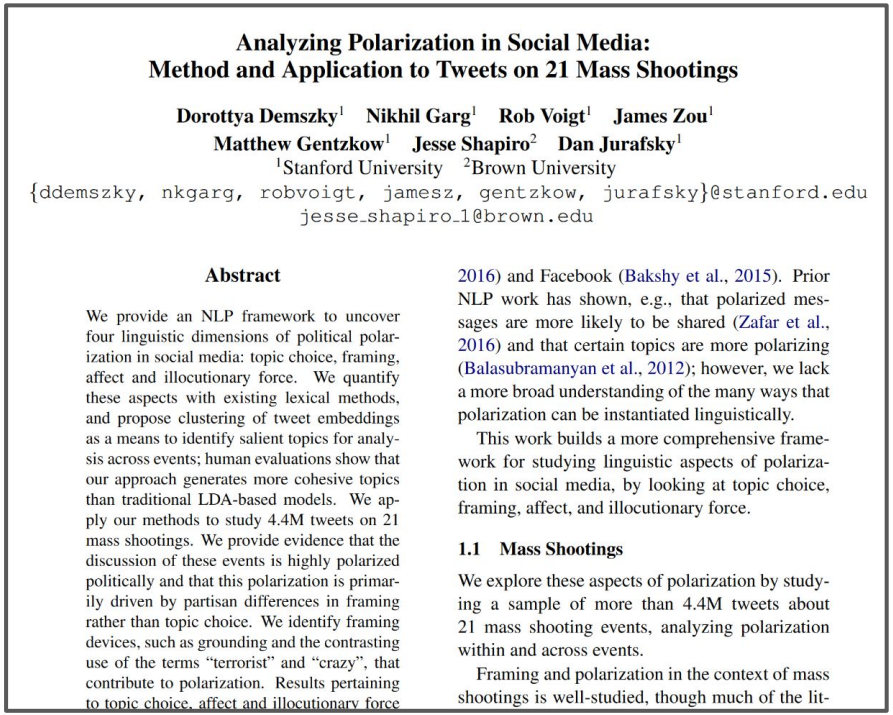
**Kristen Johnson, Di Jin, Dan Goldwasser**
Department of Computer Science
Purdue University, West Lafayette, IN 47907
{john1187, jind, dgoldwas}@purdue.edu

Johnson, Kristen, Di Jin, and Dan Goldwasser. "Modeling of political discourse framing on twitter." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. 2017.

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

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

# Topic Modeling for Social Research



**Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings**

**Dorottya Demszky**[1] **Nikhil Garg**[1] **Rob Voigt**[1] **James Zou**[1]
**Matthew Gentzkow**[1] **Jesse Shapiro**[2] **Dan Jurafsky**[1]
[1]Stanford University [2]Brown University
{ddemszky, nkgarg, robvoigt, jamesz, gentzkow, jurafsky}@stanford.edu
jesse_shapiro_1@brown.edu

**Abstract**

We provide an NLP framework to uncover four linguistic dimensions of political polarization in social media: topic choice, framing, affect and illocutionary force. We quantify these aspects with existing lexical methods, and propose clustering of tweet embeddings as a means to identify salient topics for analysis across events; human evaluations show that our approach generates more cohesive topics than traditional LDA-based models. We apply our methods to study 4.4M tweets on 21 mass shootings. We provide evidence that the discussion of these events is highly polarized politically and that this polarization is primarily driven by partisan differences in framing rather than topic choice. We identify framing devices, such as grounding and the contrasting use of the terms "terrorist" and "crazy", that contribute to polarization. Results pertaining to topic choice, affect and illocutionary force

2016) and Facebook (Bakshy et al., 2015). Prior NLP work has shown, e.g., that polarized messages are more likely to be shared (Zafar et al., 2016) and that certain topics are more polarizing (Balasubramanyan et al., 2012); however, we lack a more broad understanding of the many ways that polarization can be instantiated linguistically.

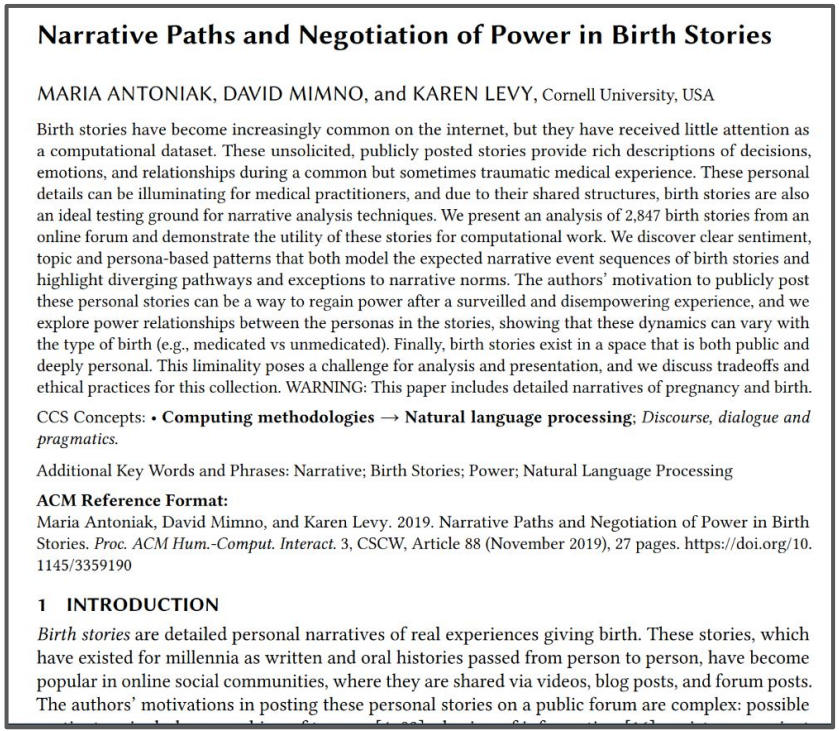
This work builds a more comprehensive framework for studying linguistic aspects of polarization in social media, by looking at topic choice, framing, affect, and illocutionary force.

### 1.1 Mass Shootings

We explore these aspects of polarization by studying a sample of more than 4.4M tweets about 21 mass shooting events, analyzing polarization within and across events.

Framing and polarization in the context of mass shootings is well-studied, though much of the lit-

[Paper](#) [Slides](#)



**Narrative Paths and Negotiation of Power in Birth Stories**

MARIA ANTONIAK, DAVID MIMNO, and KAREN LEVY, Cornell University, USA

Birth stories have become increasingly common on the internet, but they have received little attention as a computational dataset. These unsolicited, publicly posted stories provide rich descriptions of decisions, emotions, and relationships during a common but sometimes traumatic medical experience. These personal details can be illuminating for medical practitioners, and due to their shared structures, birth stories are also an ideal testing ground for narrative analysis techniques. We present an analysis of 2,847 birth stories from an online forum and demonstrate the utility of these stories for computational work. We discover clear sentiment, topic and persona-based patterns that both model the expected narrative event sequences of birth stories and highlight diverging pathways and exceptions to narrative norms. The authors' motivation to publicly post these personal stories can be a way to regain power after a surveilled and disempowering experience, and we explore power relationships between the personas in the stories, showing that these dynamics can vary with the type of birth (e.g., medicated vs unmedicated). Finally, birth stories exist in a space that is both public and deeply personal. This liminality poses a challenge for analysis and presentation, and we discuss tradeoffs and ethical practices for this collection. WARNING: This paper includes detailed narratives of pregnancy and birth.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Discourse, dialogue and pragmatics.*

Additional Key Words and Phrases: Narrative; Birth Stories; Power; Natural Language Processing

**ACM Reference Format:**
Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 88 (November 2019), 27 pages. https://doi.org/10.1145/3359190

### 1 INTRODUCTION

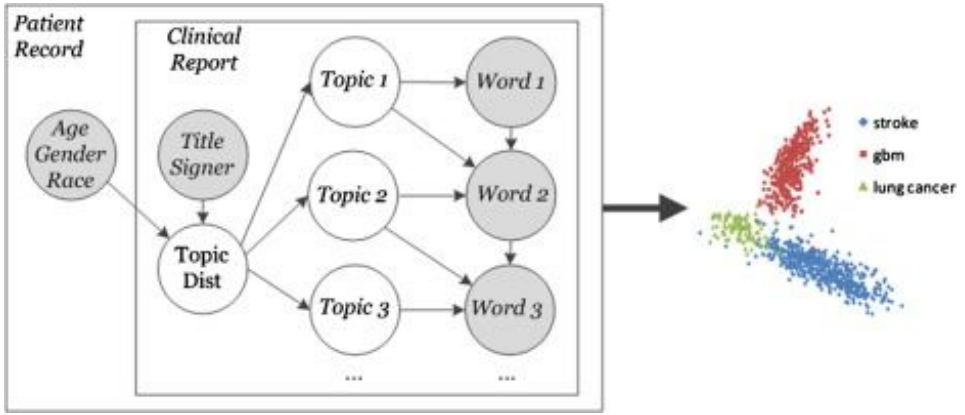
*Birth stories* are detailed personal narratives of real experiences giving birth. These stories, which have existed for millennia as written and oral histories passed from person to person, have become popular in online social communities, where they are shared via videos, blog posts, and forum posts. The authors' motivations in posting these personal stories on a public forum are complex: possible

[Paper](#) [Slides](#)

# A Little Bit of Background

All topic models are based on the same basic assumption:

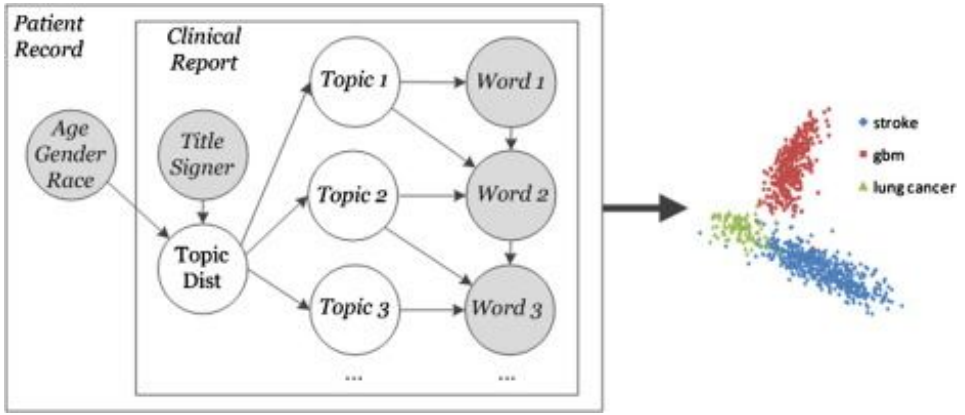
- each **document** consists of a mixture of ***topics***, and
- each **topic** consists of a collection of **words**
- The semantics of our document are governed by some hidden, or "latent" variables
- Goal of topic modeling: uncover these latent variables — topics



# A Little Bit of Background

All topic models are based on the same basic assumption:

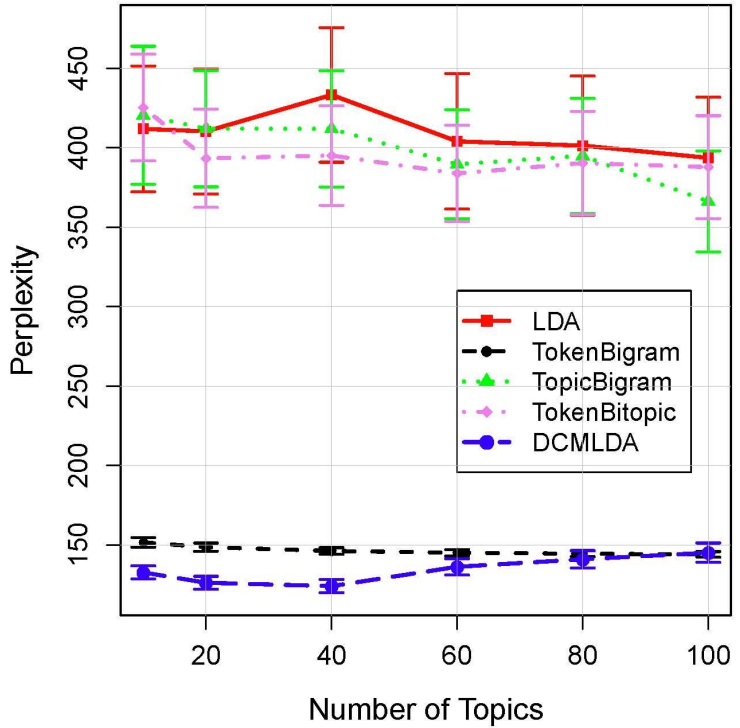
- each **document** consists of a mixture of ***topics***, and
- each **topic** consists of a collection of **words**
- The semantics of our document are governed by some hidden, or “latent” variables
- Goal of topic modeling: uncover these latent variables — topics



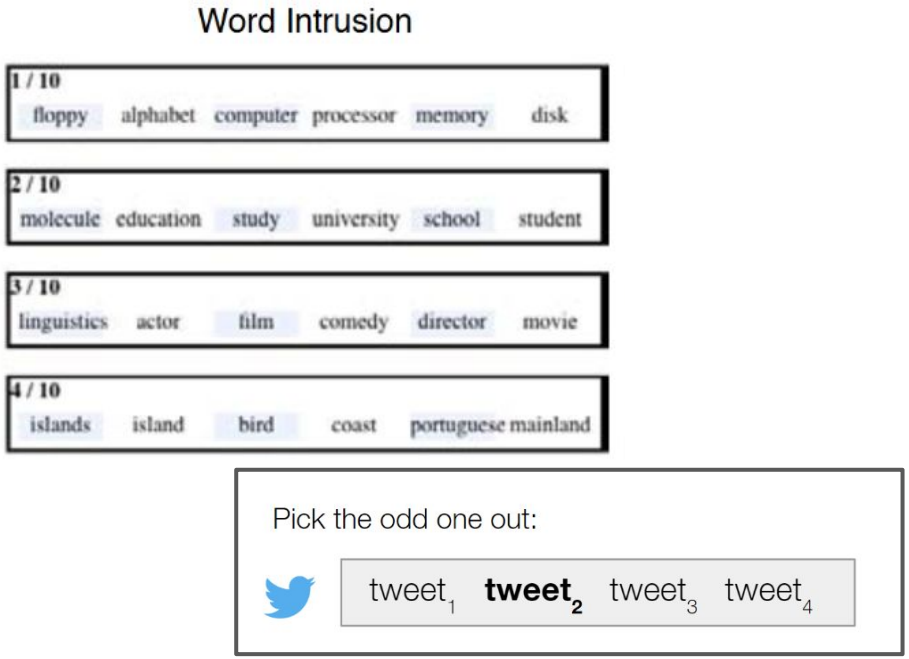
# Validating Topic Models

**Automatic / quantitative evaluation**

- **Perplexity:** widely used for picking the number of topics [e.g., https://cseweb.ucsd.edu/~elkan/perplexity.html] (the lower the perplexity, the better the fit)
- **Normalized Pointwise Mutual Information (NPMI)**
- Using LLMs (Stammbach et al., 2023): https://github.com/dominiksinsaarland/evaluating-topic-model-output



# Validating Topic Models




**Manual / qualitative evaluation**

- Direct rating (Newman et al., 2010)
- Reading 'tea-leaves': Word intrusion tests (Chang et al., 2009)
  - Evaluators see *K* words and have to identify the "most irrelevant" word — the intruder.
  - *K*-1 words chosen randomly from a topic
  - the remaining "intruder" is chosen from a different topic

- Tweet intrusion test (Demszky et al., 2019)
- oolong: An R package for validating automated content analysis tools

# Next week(s)

- **10.06: Network analysis or causal inference?**
- 17.06: Reproducible research pipeline
- 23.06: First Project Guidance Session (in-person, same room)
- 27.06: Deadline for submitting project proposal