Text Classification

Sanmay Das

Washington University in St. Louis

CSE 417T, April 18, 2019

Text Classification

- Lot of work in NLP, but we will simply treat it as a machine learning problem (ignoring parts of speech, etc).
- First question: How do we convert documents into $(\mathbf{x}, y)$ tuples?
- y depends on what we are trying to predict. Examples:
  - Sentiment (positive or negative)
  - Relevant or irrelevant to a specific product
  - Political leaning: Democrat or Republican?
  - ...
- Several different answers for $\mathbf{x}$, including
  - Representations in word / phrase space (Bag of Words, TFIDF)
  - Representations in topic space (LDA)
  - Representations in semantic space (sequences of word embeddings)

Bag of words

- Text 1: We have been put through the ringer.
- Text 2: The ringer of our telephone has been out of order.
- Text 3: A telephone is a necessity.

	we	have	been	put	through	the	ringer	of	our	telephone	has	out	order	a	is	necessity
T1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
T2	0	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0
T3	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1

Practicalities: A Real Review

This is really a new low in entertainment. Even though there are a lot worse movies out.

In the Gangster / Drug scene genre it is hard to have a convincing storyline (this movies does not, i mean Sebastians motives for example couldn't be more far fetched and worn out cliché.) Then you would also need a setting of character relationships that is believable (this movie does not.)

Sure Tristan is drawn away from his family but why was that again? what's the deal with his father again that he has to ask permission to go out at his age? ... Wasn't he already down and out, why does he do it again?

So there are some interesting questions brought up here for a solid socially critic drama (but then again, this movie is just not, because of focusing on "cool" production techniques and special effects an not giving the characters a moment to reflect and most of all forcing the story along the path where they want it to be and not paying attention to let the story breath and naturally evolve.)

It wants to be a drama to not glorify abuse of substances and violence (would be political incorrect these days, wouldn't it?) but on the other hand it is nothing more then a cheap action movie (like there are so so many out there) with an average set of actors and a Vinnie Jones who is managing to not totally ruin what's left of his reputation by doing what he always does.

So all in all i .. just ... can't recommend it.

1 for Vinnie and 2 for the editing.

Tokenization

- What constitutes a word?
- Tokenization is the process of breaking a stream of text up into meaningful “tokens” (words, phrases, symbols)
- “Wasn’t” as opposed to “Wasn”, “’t”
- “high-flying” to “high”, “flying”
- Many ways of doing this, most packages have implementations

Uninformative and Rare Words

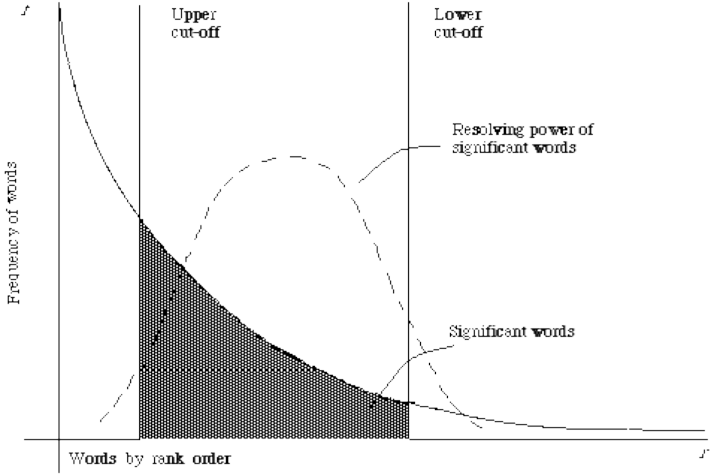


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

(Credit: Lecture notes of Hongyao Ma)

Stemming and Lemmatization

- Goal of both is to reduce forms of a word to a common base form
- am, are, is $\Rightarrow$ be
- car cars, car's, cars' $\Rightarrow$ car
- the boy's cars are different colors $\Rightarrow$ the boy car be differ color
- Stemming is a heuristic, usually quite effective. Porter stemmer is commonly used.
- Lemmatization is more complex and tries to do things properly, with linguistic analysis, etc. (may be better when using word embeddings)

(Credit: `https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html`)

N-grams and TFIDF

- N-gram representations use N words occurring in a row as the base units in $\mathbf{x}$ (blows up vocabulary, but can be more informative)
  - e.g. "personal accounts" versus "private accounts"
- TF: Term Frequency. How often a word/phrase occurs in a document (can normalize by document length in different ways)
- IDF: Document Frequency: In how many documents does this word/phrase occur? Typically use nonlinear scaling $(1 + \log(N/\textbf{doc-freq}(t)))$
  - More informative than total term frequency in the corpus

Stopword Removal

- Removing extremely common words with little value for the task
- Trend has been to move towards smaller stopword lists
  - Can break the original meaning "this is not a good movie" $\Rightarrow$ "movie"

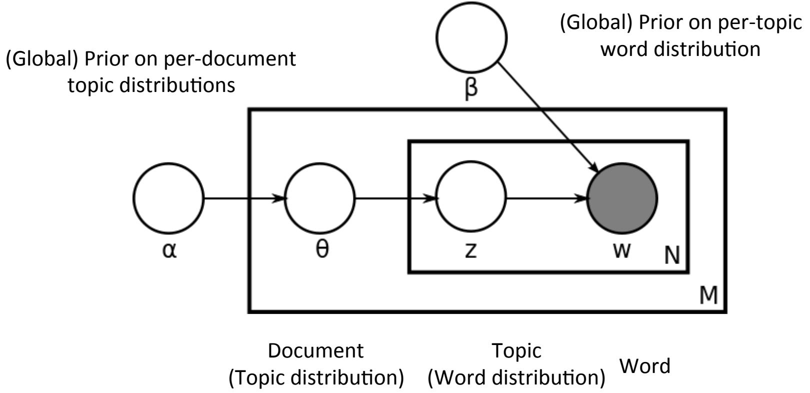
Learning Algorithms

- Typically prefer to use linear models
- Linear SVMs and regularized logistic regression (sometimes known as Maximum Entropy in the NLP / text mining literature) are quite popular

Topic Models

- Each document is a mixture of k *topics*
- Generative process: Each document is generated by repeatedly sampling:
  1. A topic from its topic distribution
  2. A word/phrase from the topic
- The topics are not necessarily semantically well-defined. They are typically based on co-occurrence patterns of words/phrases.
- The most common type of topic modeling is latent Dirichlet allocation (LDA)
  - Sparse Dirichlet prior (multivariate generalization of the Beta): encodes preference for documents coming from a smaller range of topics, and each topic being concentrated in terms of words

LDA



(Credit: By Bkkbrad - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=3610403`)

LDA Inference

- Bayesian inference problem: Find the parameters that maximize the probability of the data
  - Set of topics
  - Distribution of words for each topic
  - Topic of each word
  - Topic mixture of each document
- Typically done with Gibbs sampling, but there are other approaches

Topic Interpretation

Topics can sometimes be cleanly interpreted. For example, from some of our work on the Congressional Record (task – distinguish statements made by Republicans and Democrats):

5 topics with the highest cross validation AUC

Topic 28
(CV AUC: 0.960)
republican parti 0.007
social secur 0.006
tax cut 0.005
american peopl 0.003
wall street 0.003
great depress 0.002
liber democrat 0.002
econom polici 0.002
bill clinton 0.002
georg bush 0.002
…

Topic 36
(CV AUC: 0.950)
health care 0.021
health insur 0.006
small busi 0.006
incom tax 0.003
tax rate 0.002
tax cut 0.002
insur compani 0.002
balanc budget 0.002
million american 0.002
care system 0.002
…

Topic 11
(CV AUC: 0.945)
pro lif 0.005
onlin edit 0.004
richard dawkin 0.003
stem cell 0.002
plan parenthood 0.002
scientif medic 0.001
abort time 0.001
theori evolut 0.001
cell research 0.001
unit state 0.001
…

Topic 29
(CV AUC: 0.928)
global warm 0.006
climat chang 0.006
unit state 0.003
oil ga 0.002
natur ga 0.002
oil compani 0.002
carbon dioxid 0.002
renew energi 0.002
nuclear power 0.001
fossil fuel 0.001
…

Topic 6
(CV AUC: 0.919)
civil war 0.006
war iraq 0.003
saddam hussein 0.002
liber bia 0.002
de gaull 0.002
foreign polici 0.002
bin laden 0.002
war terror 0.002
al qaeda 0.001
middl east 0.001
…

28 Politics & the economy

36 Health care, insurance, and taxes

11 Evolutionary science, medicine

29 Climate change

6 Foreign policy in the middle east

Topic Interpretation Example (contd.)

5 topics with the lowest cross validation AUC

Topic 8
(CV AUC: 0.737)
pro choic 0.004
first amend 0.003
time limit 0.003
plan parenthood 0.002
democrat nomin 0.002
anti abort 0.002
daili show 0.002
use describ 0.002
vote bill 0.002
amend offer 0.002
…

Topic 2
(CV AUC: 0.734)
balanc time 0.004
mr speaker 0.004
breast cancer 0.003
urg colleagu 0.003
back balanc 0.002
yield back 0.002
support homeopathi 0.002
nativ american 0.002
reserv balanc 0.002
sexual assault 0.002
…

Topic 37
(CV AUC: 0.728)
talk point 0.002
rush limbaugh 0.001
hous press 0.001
liber conserv 0.001
hate group 0.001
hate speech 0.001
anti govern 0.001
club growth 0.001
donald trump 0.001
white male 0.001
…

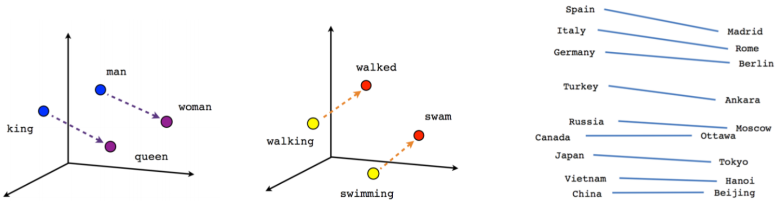
Topic 22
(CV AUC: 0.714)
new age 0.004
american polit 0.003
talk radio 0.003
parti candid 0.003
conserv movement 0.002
ayn rand 0.002
welfar state 0.002
thoma jefferson 0.002
southern state 0.002
governor new 0.001
…

Topic 3
(CV AUC: 0.683)
sarah palin 0.004
look like 0.002
new world 0.002
year ago 0.002
unit state 0.001
world order 0.001
right activist 0.001
human be 0.001
york time 0.001
mani peopl 0.001
…

8	Abortion
2	Procedural phrases
11	Opinions on media
29	Political philosophy
6	Specific people

Word Embeddings

- Sparse vocabulary vector encodings of words aren't that helpful for NNs
- Alternative: map each word into a high dimensional vector space
  - Solve the learning problem of predicting a word given context (surrounding window of words)
- Google Word2Vec has been very successful



Male-Female　　Verb tense　　Country-Capital

From https://www.tensorflow.org/versions/master/tutorials/word2vec

- Warning: Word embeddings encode human biases

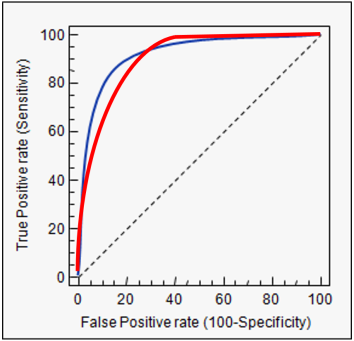
Sentiment Classification

- IMDB sentiment classification task.
- 25000 reviews each for training and testing.
- No more than 30 reviews are allowed for any given movie.
- Train and test sets contain a disjoint set of movies.
- Negative review: score ≤ 4 (out of 10)
- Positive review: score ≥ 7.
- Current methods achieve accuracy well above 90%

A Cautionary Tale: Political Partisanship / Ideology Measurement

- Idea: Measure ideology by building a classifier / regression model of partisanship
- Main datasets:
  - Congressional Record and press releases (politicians)
  - Salon and Townhall (media)
  - Conservapedia, RationalWiki (the crowd)
- Algorithms:
  - Logistic regression on n-grams (Bag-of-bigrams, TFIDF, feature hashing) + domain adaptation when needed (mSDA)
  - Recursive autoencoder (RAE).
- Labels:
  - Classification target: “Democrat” vs. “Republican”
  - Regression target: DW-nominate score (measure of ideology based on roll-call votes)

ROC Curves

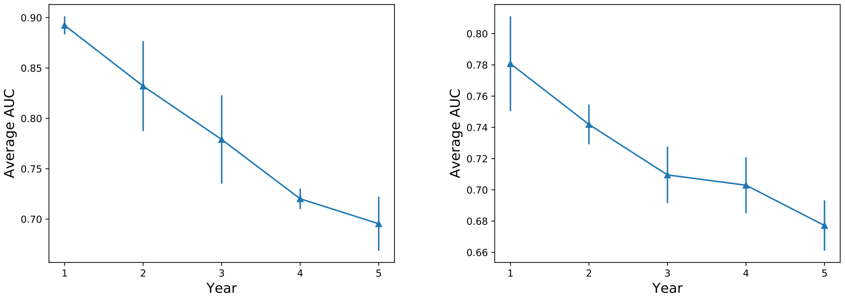


- Useful summary metric: Area Under the ROC Curve (AUC)
  - Probability a random positive example is ranked higher than a random negative example
  - Insensitive to class imbalance

(Credit: `https://stats.stackexchange.com/questions/264477/will-roc-curve-for-a-model-always-be-symmetric-if-we-have-enough-training-data`)

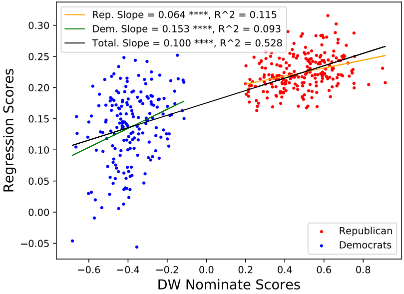
Caution 1: Generalizing Across Time

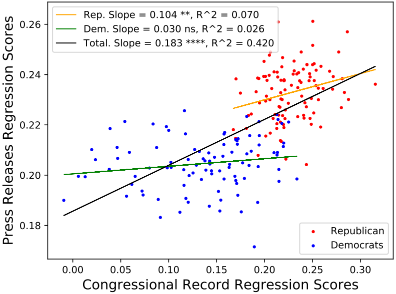
Out-of-time predictions for Salon-Townhall (left) and the Congressional Record (right) when trained on two years of data and tested going forward.



Caution 2: Extrapolating From Partisanship to Ideology

Scatterplot of ideology predictions based on the Congressional Record vs. DW-Nominate scores (left), and of ideology predictions based on the Congressional Record vs. based on press releases (right) for house members in the 113th Congress (2013-2014).





Caution 3: Generalizing Across Datasets

Testing and Averaging By Year

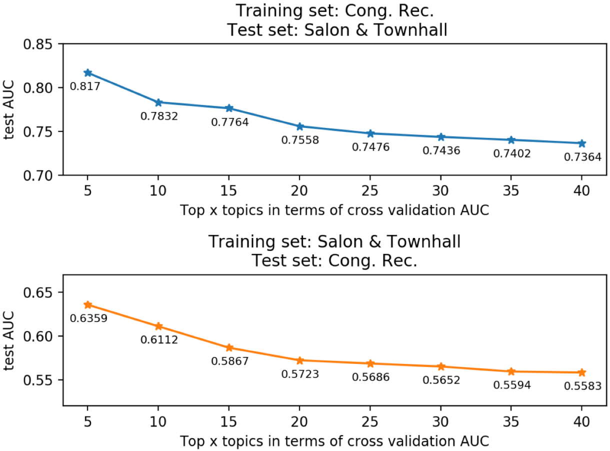
Training \ Test	Congressional Record	Salon & Townhall	Conservapedia & RationalWiki
Congressional Record	0.83 (TF-IDFLR) 0.81 (RAE)	**0.69**(mSDA) 0.67 (TF-IDFLR) 0.59(RAE)	0.47(mSDA) **0.49** (TF-IDFLR) 0.47 (RAE)
Salon & Townhall	**0.60**(mSDA) 0.59 (TF-IDFLR) 0.54 (RAE)	0.92(TF-IDFLR) 0.90(RAE)	0.52(mSDA) 0.51 (TF-IDFLR) **0.55**(RAE)
Conservapedia & RationalWiki	**0.53**(mSDA) 0.50 (TF-IDFLR) 0.47 (RAE)	**0.58**(mSDA) 0.53 (TF-IDFLR) 0.57 (RAE)	0.85 (TF-IDFLR) 0.82 (RAE)

A Silver Lining?

- Predictability may be a function of topic
- Learn a topic model jointly on CR and ST
- Learn individual classifiers for texts “hard classified” to each of 40 topics

A Silver Lining?

- Predictability may be a function of topic
- Learn a topic model jointly on CR and ST
- Learn individual classifiers for texts “hard classified” to each of 40 topics



High AUC Topics

5 topics with the highest cross validation AUC

Topic 28
(CV AUC: 0.960)
republican parti 0.007
social secur 0.006
tax cut 0.005
american peopl 0.003
wall street 0.003
great depress 0.002
liber democrat 0.002
econom polici 0.002
bill clinton 0.002
georg bush 0.002
…

Topic 36
(CV AUC: 0.950)
health care 0.021
health insur 0.006
small busi 0.006
incom tax 0.003
tax rate 0.002
tax cut 0.002
insur compani 0.002
balanc budget 0.002
million american 0.002
care system 0.002
…

Topic 11
(CV AUC: 0.945)
pro lif 0.005
onlin edit 0.004
richard dawkin 0.003
stem cell 0.002
plan parenthood 0.002
scientif medic 0.001
abort time 0.001
theori evolut 0.001
cell research 0.001
unit state 0.001
…

Topic 29
(CV AUC: 0.928)
global warm 0.006
climat chang 0.006
unit state 0.003
oil ga 0.002
natur ga 0.002
oil compani 0.002
carbon dioxid 0.002
renew energi 0.002
nuclear power 0.001
fossil fuel 0.001
…

Topic 6
(CV AUC: 0.919)
civil war 0.006
war iraq 0.003
saddam hussein 0.002
liber bia 0.002
de gaull 0.002
foreign polici 0.002
bin laden 0.002
war terror 0.002
al qaeda 0.001
middl east 0.001
…

28 Politics & the economy
36 Health care, insurance, and taxes
11 Evolutionary science, medicine
29 Climate change
6 Foreign policy in the middle east

Low AUC Topics

5 topics with the lowest cross validation AUC

Topic 8
(CV AUC: 0.737)
pro choic 0.004
first amend 0.003
time limit 0.003
plan parenthood 0.002
democrat nomin 0.002
anti abort 0.002
daili show 0.002
use describ 0.002
vote bill 0.002
amend offer 0.002
…

Topic 2
(CV AUC: 0.734)
balanc time 0.004
mr speaker 0.004
breast cancer 0.003
urg colleagu 0.003
back balanc 0.002
yield back 0.002
support homeopathi 0.002
nativ american 0.002
reserv balanc 0.002
sexual assault 0.002
…

Topic 37
(CV AUC: 0.728)
talk point 0.002
rush limbaugh 0.001
hous press 0.001
liber conserv 0.001
hate group 0.001
hate speech 0.001
anti govern 0.001
club growth 0.001
donald trump 0.001
white male 0.001
…

Topic 22
(CV AUC: 0.714)
new age 0.004
american polit 0.003
talk radio 0.003
parti candid 0.003
conserv movement 0.002
ayn rand 0.002
welfar state 0.002
thoma jefferson 0.002
southern state 0.002
governor new 0.001
…

Topic 3
(CV AUC: 0.683)
sarah palin 0.004
look like 0.002
new world 0.002
year ago 0.002
unit state 0.001
world order 0.001
right activist 0.001
human be 0.001
york time 0.001
mani peopl 0.001
…

8	Abortion
2	Procedural phrases
11	Opinions on media
29	Political philosophy
6	Specific people