Quantitative text analysis: Current topics

Blake Miller

MY 459: Quantitative Text Analysis

March 27, 2023

Course website: lse-my459.github.io

1. Overview and Fundamentals
2. Descriptive Statistical Methods for Text Analysis
3. Automated Dictionary Methods
4. Machine Learning for Texts
5. Supervised Scaling Models for Texts
6. *Reading Week*
7. Unsupervised Models for Scaling Texts
8. Similarity and Clustering Methods
9. Topic models
10. Word embeddings
11. Current topics

Today

- Beyond the bag of words
- Considering biases in social media data
- Guided coding

- Beyond the bag of words
- Considering biases in social media data
- Guided coding

Why Bag of Words?

- In the course, we focused on bag of words (B.O.W.) models because they are considered the best choice for a wide range of datasets and tasks in text analysis in the social sciences.
- B.O.W. classifiers that are based on term frequencies have shown to achieve good performance on many tasks and are of particular benefit to social scientists because they are easy to interpret.

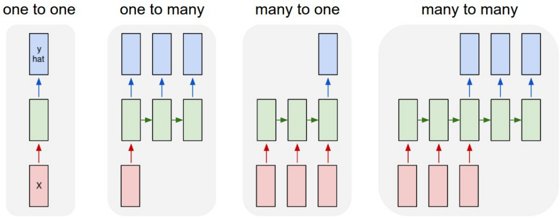
Limitations of Bag of Words

- When the goal is prediction rather than inference, more complex models of language can be more beneficial; we throw out a lot of information with the B.O.W. assumption.
- The interdependent nature of words becomes important in many natural language processing (NLP) tasks, and understanding the relationships between words is critical for accurate analysis.
- While B.O.W. models can be effective, they struggle with complexities of language (e.g., emotions, irony, sarcasm).
- The following slides mention a few common models to address these complexities and provide links to further materials should you wish to study these topics more in the future.

Recurrent neural networks

- Recurrent neural networks (RNNs) are one example of models that can capture dependencies between words in language
- They can process sequences of inputs and predict sequences of outputs (not restricted to words/language)
- RNNs are used in a range of tasks in natural language processing, e.g. classification, image captioning, or machine translation
- Most common types of RNNs such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) are based on cells which improve the model's ability to remember long term dependencies

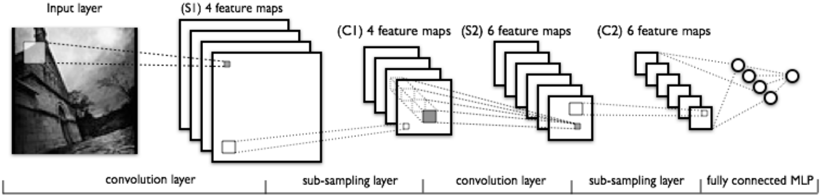
Recurrent neural networks



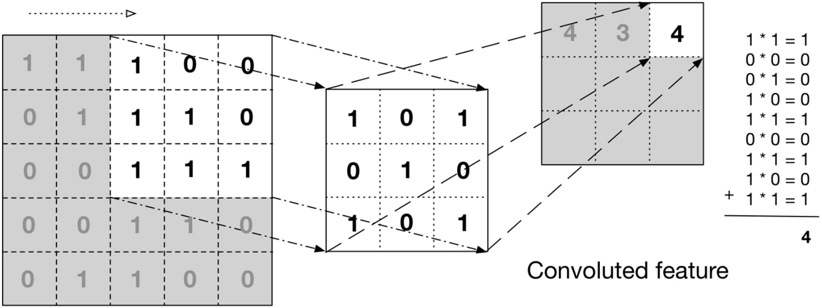
Source: From Andrej Karpathy's blog; slightly edited

- Arrows are functions/transformations, rectangles are vectors, green rectangles hold states
- One to one: Standard feed forward neural network
- One to many: RNN that e.g. takes an image as input and then outputs a sentence describing it
- Many to one: RNN that e.g. inputs a sequence of words and outputs a sentiment label
- Many to many: RNN that e.g. inputs a sentence in one language and outputs it in another language

Convolutional neural networks for images



Convolutional neural networks for images



1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Input data

1	0	1
0	1	0
1	0	1

Kernel

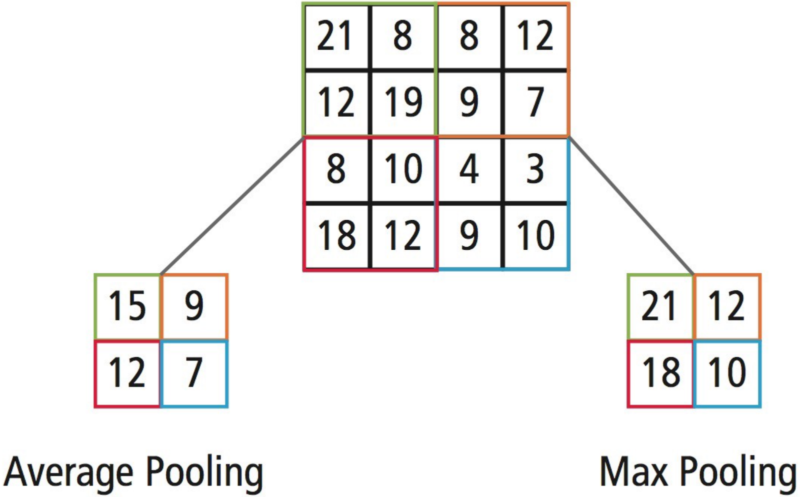
4	3	4

Convoluted feature

1 * 1 = 1
0 * 0 = 0
0 * 1 = 0
1 * 0 = 0
1 * 1 = 1
0 * 0 = 0
1 * 1 = 1
1 * 0 = 0
+ 1 * 1 = 1

4

Convolutional neural networks for images

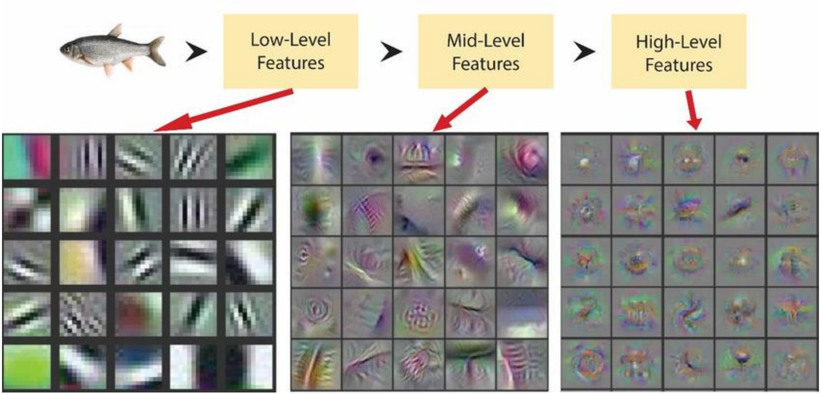


21	8	8	12
12	19	9	7
8	10	4	3
18	12	9	10

15	9
12	7

21	12
18	10

Convolutional neural networks for images

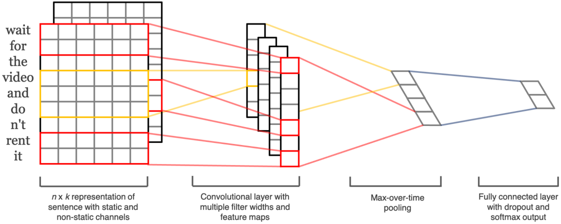


Convolutional neural networks for language

- Also convolutional neural networks (CNNs), originally from computer vision, can take the order of words into account

- The following model e.g. achieves very good performance in the classification of short sentences

- Word embeddings of words in a sentence are arranged like an "image" and hence make it possible to use this model from computer vision for sentences


wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Source: Kim (2014)

Convolutional neural networks for language

- Why do they improve over bag of words models?
  - Term dependencies are captured (even long distance dependencies) where bag of words models ignore dependencies.
  - Word context is considered explicitly through convolutions and embeddings.
  - Use of embeddings automatically creates equivalence classes between words (e.g., 'good' and 'great' will have very similar vectors)
  - Bag of words models are often not so great at classifying short texts (lots of information thrown out through bag of words assumption)
- For an explainer on CNNs in more detail, a good place to start is here

Transformers

- Newer models are e.g. transformers which are very frequently used e.g. in machine translation today (Vaswani et al. 2017, https://arxiv.org/abs/1706.03762)
- Their architecture features an encoder and a decoder
- The encoder transforms a set of input words *simultaneously* into embeddings that represent their meaning in the original language
- The decoder then uses these embeddings to predict the associated words in the other language
- So call "attention" is a key feature of these models. Rather than sequentially, the models process a set of words all at once and then direct attention to words selectively
- Their architecture favours parallelisation, which decreases the time necessary to train them

BERT

- A very popular transformer based model in the last couple of years has been BERT (Devlin et al. 2018, https://arxiv.org/abs/1810.04805)
- This model stacks transformer encoders and is able to produce exceptionally good word embeddings when sets of words such as sentences are parsed into it
- The BERT model can be downloaded pre-trained and adapted to a range of tasks
- In sentiment classification, for example, mainly an added function between the embeddings and the sentiment labels is learned
- This much decreases the time necessary to train the model

Further study: Deep learning and natural language processing

- Should you wish to study deep learning and natural language processing in the future, the following course is freely available online http://web.stanford.edu/class/cs224n/ (the last publicly available videos correspond to the course version from 2021 and can be found here)
- The course uses Python which is the more common language for neural networks and deep learning
- To implement neural networks in R, see e.g. these Tensorflow/Keras tutorials. The following repo contains a range of baseline code examples for Keras neural network implementations in R

- Beyond the bag of words
- Considering biases in social media data
- Guided coding

Biases in sampling

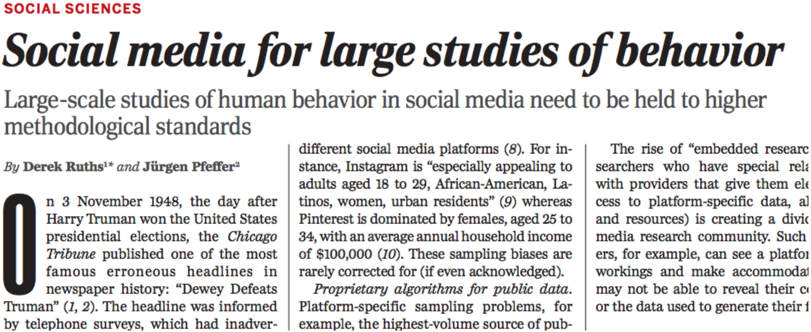
Morstatter et al, 2013, *ICWSM*, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose":

- 1% random sample from Twitter's Streaming API is not truly random
- Less popular hashtags, users, topics... less likely to be sampled
- But for keyword-based samples, bias is not as important

González-Bailón et al, 2014, *Social Networks*, "Assessing the bias in samples of large online networks":

- Small samples collected by filtering with a subset of relevant hashtags can be biased
- Central, most active users are more likely to be sampled
- Data collected via search (REST) API more biased than those collected with Streaming API

Biases in social media data more general



SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By Derek Ruths[1]* and Jürgen Pfeffer[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: “Dewey Defeats Truman” (*1*, *2*). The headline was informed by telephone surveys, which had inadver-

different social media platforms (*8*). For instance, Instagram is “especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents” (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

Proprietary algorithms for public data. Platform-specific sampling problems, for example, the highest-volume source of pub-

The rise of “embedded researc
searchers who have special rela
with providers that give them ele
cess to platform-specific data, al
and resources) is creating a divic
media research community. Such
ers, for example, can see a platfor
workings and make accommodat
may not be able to reveal their c
or the data used to generate their f

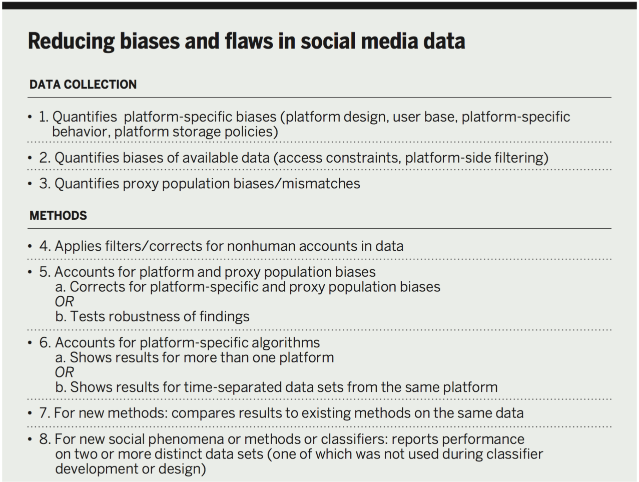
Ruths and Pfeffer, 2015, “Social media for large studies of behavior”, *Science*

Biases in social media data more general

Sources of bias (Ruths and Pfeffer, 2015; Lazer et al, 2017)

- Population bias
  - Sociodemographic characteristics are correlated with presence on social media
- Self-selection within samples
  - Partisans more likely to post about politics (Barberá & Rivero, 2014)
- Proprietary algorithms for public data
  - Twitter API does not always return 100% of publicly available tweets (Morstatter et al, 2014)
- Human behavior and online platform design
  - e.g. *Google Flu* (Lazer et al, 2014)

Biases in social media data more general



Reducing biases and flaws in social media data

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
  a. Corrects for platform-specific and proxy population biases
  OR
  b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
  a. Shows results for more than one platform
  OR
  b. Shows results for time-separated data sets from the same platform
- 7. For new methods: compares results to existing methods on the same data
- 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct data sets (one of which was not used during classifier development or design)

Issues in evaluating data from social media. Large-scale social media studies of human behavior should i address issues listed and discussed herein (further discussion in supplementary materials).

Ruths and Pfeffer, 2015, “Social media for large studies of behavior”, *Science*

- Beyond the bag of words
- Considering biases in social media data
- Guided coding

Guided coding

- Today we are going to look at case study about building a machine learning classifier that tries to predict whether a sentence is violent and briefly demonstrate how one would use active learning to aid in the process of building a training set.
- We will then demonstrate how one might classify “personal attacks” in social media using some models we learned in class and then I’ll briefly demonstrate how we might go about classifying personal attacks using a convolutional neural network (CNN) for text in Python.

Guided coding

- 01-classifying-violent-speech.Rmd
- 02-toxic-comments.Rmd
- 03-toxic-comments-CNN.py