Quantitative text analysis: Current topics

Friedrich Geiecke

MY 459: Quantitative Text Analysis

March 28, 2022

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
- The Twitter API and social media data
- Guided coding

- Beyond the bag of words
- The Twitter API and social media data
- Guided coding

Demo

- Let us begin with a demo of a very recent model to detect emotions in texts

How could this work?

- Recent developments in AI are often driven by machine learning
- When seeing impressive results like this, a first step is therefore to think about the broad categories which characterise machine learning:
  - 1. Supervised learning: Learning the function between X and y
  - 2. Unsupervised learning: Learning patterns in X
  - 3. Reinforcement learning: Solving dynamic problems

How could this work?

- The setup suggests that this could be supervised learning model: A sentence (x) predicts an emotion label y
- The difficult function between X and y suggests it has has to be a very flexible model:
- "Had a great day" needs to result in an entirely different prediction than "Had a great day ... not"
- Furthermore, input words in such models might be represented as word embeddings obtained from unsupervised learning

Answer

- The model is from the DeepMoji project https://deepmoji.mit.edu/ by Felbo et al. (2017)
- A deep neural network was trained on around 1.2 billion tweets
- Each tweet contains one of 64 common emojis
- The emojis are separated from the text and the model simply predicts the emoji (y) from the tweet text (x), but does this very well
- Applied to a new text without emojis, the model predicts suitable emojis
- In the demo I grouped emojis into broad categories and only reported the categories

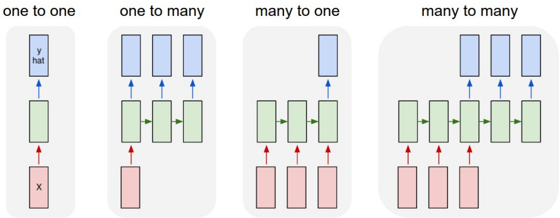
Capturing dependencies in language

- In the course we focussed on bag of words models because they are the best choice for a wide range of datasets and tasks in text analysis in the social sciences
- Bag of word classifiers based on term frequencies can also classify tweets into emojis and achieve good performance
- Yet, when the interdependent nature of words becomes as important as in the case of detecting emotions, irony, etc. more advanced models can become helpful that capture the dependencies in language
- The following slides mention a few common types of models 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

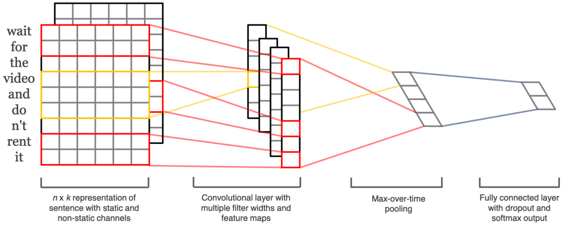


Source: From Andrej Karpathy's blog; slightly edited

- 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 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



Source: Kim (2014)

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

This lecture

- We will now continue with a discussion of the Twitter API and Twitter data as an application of social media data
- On the one hand tweets are an important example of social media data that is frequently studied by social scientists today
- On the other hand using the Twitter API will eventually allow us to connect all the dots in the coding session where we will try to develop a classifier which approximates whether a sentence agrees or disagrees

- Beyond the bag of words
- The Twitter API and social media data
- Guided coding

APIs

- API: Application Programming Interface
- In web APIs, a set of structured HTTP requests can return data in a lightweight format e.g. JSON or XML
- The API user sends a request to the API (e.g. with a software such as R) and the API returns data from the API provider's database
- We will use the 'rtweet' package to to access the Twitter API from R

Why APIs?

Advantages

- Cleaner data collection: Avoid malformed HTML, no legal issues, clear data structures, more trust in data collection...
- Standardized data access procedures: Transparency, replicability
- Robustness: Benefits from "wisdom of the crowds"

Disadvantages

- Not always available
- Dependency on API providers
- Rate limits

Twitter APIs

Two different methods to collect Twitter data

1. REST API
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: User profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
2. Streaming API
   - Connect to the "stream" of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Sample stream: 1% random sample of tweets
     - 2.2 Filter stream: tweets filtered by keywords (when volume reaches 1% of all tweets, it will also return a random sample)
     - 2.3 Geo stream: tweets filtered by location

Twitter APIs

- Tweets can only be downloaded in real time, historical data is generally much harder to obtain (exceptions: last seven days or user timelines, where $\sim$ 3,200 most recent tweets are available)
- Very recent special access for researchers allows to obtain more historical data

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 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*

Addendum: Academic Research and the Twitter API

- Very recently, an "Academic Research product track" for the Twitter API was introduced
- Among other features, it can be used to access significant amounts of historical tweets for free if the application is approved
- Applications can be made via `https://developer.twitter.com/en/products/twitter-api/academic-research`
- Non-commercial use only and requires a clearly defined research objective
- There also exists an R package specifically for this type of API **academictwitteR**

- Beyond the bag of words
- The Twitter API and 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 might contain approval or disapproval
- For this we will go through the process of building such a model step by step, from the data collection to training

Guided coding

- 01-streaming-tweets.Rmd
- 02-pre-processing.Rmd
- 03-tf-classifiers.Rmd
- 04-avg-embedding-classifier.Rmd
- 05-deep-classifier.Rmd