STCS 6701: Foundations of Graphical Models: Reading 7

Maxime TCHIBOZO (MT3390)

October 2020

1 Semantics Derived Automatically From Language Corpora Contain Human-Like Biases (2017) - Caliskan et al.

While not directly related to graphical models, this paper is relevant for my final project which will study bias & ambiguity in text data and sentiment analysis.

In Machine Learning, bias refers to prior information necessary for intelligent action. Bias can be problematic when the prior information is derived from aspects of human culture known to lead to harmful behavior (stereotyped biases lead to prejudiced actions). If these biases are unaccounted for, cultural stereotypes propagate to artificial intelligence technologies.

The Implicit Association Test measures a spectrum of known human biases in text corpuses using purely statistical methods. IAT measures response time when subjects are asked to pair two concepts they find similar, in contrast to two concepts they find different. For example: IAT finds that flowers are more pleasant than insects by comparing the latencies of four pairings (flowers + pleasant, insects + unpleasant, flowers + unpleasant and insects + pleasant. IAT is widely used in social science.

The authors of this paper developed the Word-Embedding Association Test (WEAT), a statistical test similar to IAT but which works on semantic representations of words (GloVe word embeddings) instead of words directly, with the distance between pairs of vectors (cosine similarity) analogous to reaction time in IAT.

The authors find that word embeddings – which learn properties of the object they represent using only the surrounding words – absorb stereotypical bias just like words do. Pleasant-ness association tests find European-American name embeddings to be more associated to pleasant-ness and female name embeddings more associated to family than career words. Female name embeddings are more associated with the arts than with the sciences when compared to male name embeddings – this matches IAT findings over words.

The paper then analyses the following question: do the gender-occupation word embedding biases correlate with the true gender distributions in the labor-force. To answer this question, researchers develop the Word-Embedding Factual Association Test (WEFAT), and find GloVe word embedding biases to correlate strongly with the labor occupation of women in the United States. A notable caveat, is that these findings are specific to the used English language corpus, and closely matched the approach used by the IAT.

The authors conclude that caution must be used in incorporating modules constructed via unsupervised Machine Learning into decision-making systems. This is typically something one could consider when developing Latent Dirichlet Allocation topics, or Bayesian Recommendation systems reliant on word embeddings.

Short paper, but by far one of the most fascinating I have ever read.

Follow up readings:
Big Data's Disparate Impact (2014) - Barocas and Selbst

Word embeddings can perpetuate cultural stereotypes. If resume screening is subject to cultural stereotypes, it will result in prejudiced outcomes.

Harvesting implicit group attitudes and beliefs from a demonstration website (2002) - Nosek et al.

We can measure implicit bias through web experiments. User responses demonstrate on average an implicit preference for White over Black and young over old, with other stereotypic associations such as linking male terms with science and career and female terms with liberal arts and family.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (2016) - Bolukbasi et al.

Defines metrics to quantify both direct and indirect gender biases in word embeddings, and develops algorithms to "debias" the embeddings.

Fairness Through Awareness (2011) - Dwork et al.

Authors define "fair affirmative action" which guarantees statistical parity (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), as a possible metric to evaluate bias in algorithms.

Certifying and removing disparate impact (2014) - Feldman et al.

Defines bias in algorithms through the legal notion of disparate impact (the same selection process has different outcomes for different groups, even if it appears to be neutral). The authors provide classification accuracy metrics to evaluate disparate impact, a test measuring information leakage for a hidden class (e.g. candidate's race, gender) from other data attributes, and possible approaches to make algorithms unbiased without losing relevant information in the data.

2 VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (2014) – Hutto and Gilbert

VADER is a dictionary which associates a polarity score (positive/negative/neutral sentiment) and valence score (intensity of the sentiment) to words. The authors of this paper refer to such dictionaries as "lexicons". This paper reviews existing lexicons, highlights their limitations, and indicates how VADER alleviates some of these limitations.

Most lexicons are "gold standard lexicons", meaning they have been manually created by human interpreters who assign valence (usually nuanced, continuous scores) or polarity (usually binary scores) scores to words using expert domain knowledge. Such manually created lexicons include Linguistic Inquiry and Word Count (LIWC – dictionary of 406 positive emotion words, and 499 negative emotion words; binary scores), General Inquirer (GI - 11k words across 183 categories, including 1,915 positive words and 2,291 negative words; binary scores) and Affective Norms for English Words (ANEW – words with emotional ratings: 1,034 words ranked in terms of pleasure, arousal and dominance; integer scores).

SentiWordNet (147,306 words with a positive, neutral and negative score; continuous score) uses semi-supervised learning to assign scores to words, and SenticNet (14,244 words with a polarity ranging from -1 to 1; continuous score) uses graph mining and dimensionality reduction for data driven scoring. Other data-driven approaches include Naïve Bayes classification, Maximum Entropy (i.e. logistic regression), and SVM, but the authors of the paper indicate that these require extensive training data and are computationally expensive to train.

The authors of the paper claim that all of the aforementioned lexicons do not correctly account for the intensity of a sentiment, specifically when this intensity is conveyed through punctuation or emojis. This makes the previous lexicons unsuited to blogs, tweets and other social media text data.

The VADER dictionary was created a gold standard sentiment lexicon sensitive to both polarity and intensity of the sentiment. Each word is assigned polarity and intensity scores, and the authors use simple rule-based algorithms to determine the polarity of each sentence. This approach is essentially human-centered, uses data from LIWC, ANEW and GI, and augments this data by adding in social media-specific lexicons (emoticons, acronyms, slang). Grounded Theory is then used to estable heuristics on how punctuation, emojis, capital letters and the lexicon interact together to assess a sentiment score from text. The model is evaluated using ground truth human-rated documents, and the results are compared to LIWC, GI, ANEW, Hu-Liu04, WSD, SentiWordNet, SenticNet, Naïve Bayes, Maximum Entropy and SVM (classification and regression).

In their conclusion, the authors mention promising work on Recursive Deep Learning models for sentence-level sentiment analyis by Socher et al. (2013) which achieved state of the art in several NLP categories.