Lexicons on Demand: Neural Word Embeddings for Large-Scale Text Analysis

Ethan Fast, Binbin Chen, Michael S. Bernstein
Stanford University
{ethanfast, msb}@cs.stanford.edu, bchen42@stanford.edu

Abstract

Human language is colored by a broad range of topics, but existing text analysis tools only focus on a small number of them. We present *Empath*, a tool that can generate and validate new lexical categories on demand from a small set of seed terms (like "bleed" and "punch" to generate the category *violence*). Empath draws connotations between words and phrases by learning a neural embedding across billions of words on the web. Given a small set of seed words that characterize a category, Empath uses its neural embedding to discover new related terms, then validates the category with a crowd-powered filter. Empath also analyzes text across 200 built-in, pre-validated categories we have generated such as *neglect*, *government*, and *social media*. We show that Empath's data-driven, human validated categories are highly correlated (r=0.906) with similar categories in LIWC.

1 Introduction

Language is rich in subtle signals. The previous sentence, for example, conveys connotations of *wealth* ("rich"), *cleverness* ("subtle"), *communication* ("language", "signals"), and *positive sentiment* ("rich"). A growing body of work in human-computer interaction, computational social science and social computing uses tools to identify these signals: for example, detecting emotional contagion in status updates or linguistic correlates of deception [Kramer *et al.*, 2014; Ott *et al.*, 2011]. As we gain access to ever larger and more diverse datasets, it becomes important to scale our ability to conduct such analyses with breadth and accuracy.

High quality lexicons allow us to analyze language at scale and across a broad range of signals. For example, researchers often use LIWC (Linguistic Inquiry and Word Count) to analyze social media posts, counting words in lexical categories like *sadness*, *health*, and *positive emotion* [Pennebaker *et al.*, 2001]. LIWC offers many advantages: it is fast, easy to interpret, and extensively validated. Researchers can easily inspect and modify the terms in its categories — word lists that, for example, relate "scream" and "war" to the emotion *anger*. But like other popular lexicons, LIWC is small: it has only 40 topical and emotional categories, many of which contain fewer than 100 words. Further, many potentially useful categories like *violence* or *social media* don't exist in current lexicons, requiring creating of new gold standard word lists. Other categories may benefit from updating with modern terms like "paypal" for *money* or "selfie" for *leisure*.

We present *Empath*: a lexicon mined from modern text on the web. Empath allows researchers to generate and validate new lexical categories on demand, using a combination of machine learning and crowdsourcing. For example, using the seed terms "twitter" and "facebook," we can generate and validate a category for *social media*. Empath also analyzes text across 200 built-in, pre-validated categories such as *neglect* (deprive, refusal), *government* (embassy, democrat), *strength* (tough, forceful), and *technology* (ipad, android). Empath combines modern NLP techniques with the benefits of handmade lexicons: its categories are word lists, easily extended and fast. And like LIWC (but unlike other machine learning models), Empath's contents are validated by humans.

To build Empath, we apply a skip-gram network to capture words in a neural embedding [Mikolov *et al.*, 2013a]. This embedding learns associations between words and their context, providing a model of connotation. We can then use similarity comparisons in the resulting vector space to map a vocabulary of 59,690 words onto Empath's 200 categories (and beyond, onto user-defined categories). Finally, we demonstrate how we can filter these relationships through the crowd to efficiently construct new, human validated dictionaries.

In evaluation, we show how Empath's model can replicate and extend classic work in classifying deceptive language [Ott *et al.*, 2011] and analyzing mood on twitter [Golder and Macy, 2011]. We then further validate Empath by comparing its analyses against LIWC, a lexicon of gold standard categories that have been psychometrically validated. We find the correlation between Empath and LIWC across a mixed-corpus dataset is high both with (r=0.906) and without (0.90) the crowd filter. In sum, Empath shares high correlation with gold standard lexicons, yet it also offers analyses over a dynamic set of categories.

2 Related Work

Empath inherits from a rich ecosystem of tools and applications for text analysis, and draws on the insights of prior work in data mining and unsupervised language modeling.

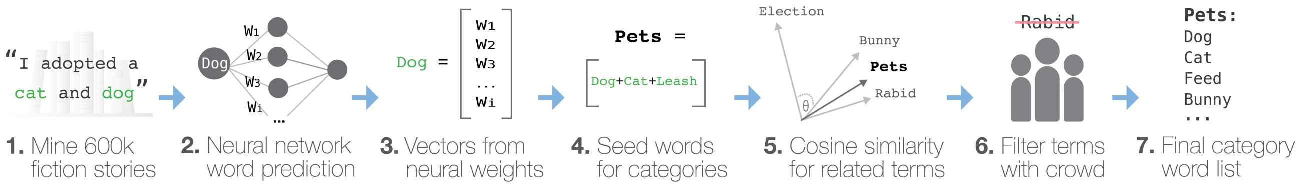
"I adopted a cat and dog"
Dog
W1
W2
W3
Wi
...
Dog = [W1 W2 W3 ... Wi]
Pets =
[Dog+Cat+Leash]
Election
Bunny
Pets
Rabid
θ
~~Rabid~~
Pets:
Dog
Cat
Feed
Bunny
...
1. Mine 600k fiction stories
2. Neural network word prediction
3. Vectors from neural weights
4. Seed words for categories
5. Cosine similarity for related terms
6. Filter terms with crowd
7. Final category word list

Figure 1: Empath learns word embeddings from 1.8 billion words of fiction, makes a vector space from these embeddings that measures the similarity between words, uses seed terms to define and discover new words for each of its categories, and finally filters its categories using crowds.

2.1 Extracting signal from text

Text analysis via dictionary categories has a long history in academic research. LIWC, for example, is an extensively validated dictionary that offers a total of 62 syntactic (e.g., present tense verbs, pronouns), topical (e.g., home, work, family) and emotional (e.g., anger, sadness) categories [Pennebaker *et al.*, 2001]. The General Inquirer (GI) is another human curated dictionary that operates over a broader set of topics than LIWC (e.g., power, weakness), but fewer emotions [Stone *et al.*, 1966]. Other tools like EmoLex, ANEW, and SentiWordNet are designed to analyze larger sets of emotional categories [Mohammad and Turney, 2013; Bradley and Lang, 1999; Esuli and Sebastiani, 2006]. While Empath's analyses are similarly driven by dictionary-based word counts, Empath operates over a more extensive set of categories, and can generate and validate new categories on demand using unsupervised language modeling.

Work in sentiment analysis has developed powerful techniques to classify text across positive and negative polarity [Socher *et al.*, 2013], but has also benefited from simpler, transparent models and rules [Hutto and Gilbert, 2014]. Empath draws on the complementary strengths of these ideas, using the power of unsupervised machine learning to create *human-interpretable* feature sets for the analysis of text. One of Empath's goals is to embed modern NLP techniques in a way that offers the transparency of dictionaries like LIWC.

2.2 Data mining and modeling

A large body of prior work has investigated unsupervised language modeling. For example, researchers have learned sentiment models from the relationships between words [Hatzivassiloglou and McKeown, 1997], classified the polarity of reviews in an unsupervised fashion [Turney, 2002], discovered patterns of narrative in text [Chambers and Jurafsky, 2009], and (more recently) used neural networks to model word meanings in a vector space [Mikolov *et al.*, 2013a]. We borrow from the last of these approaches in constructing of Empath's unsupervised model.

Empath also takes inspiration from techniques for mining human patterns from data. Augur likewise mines text on the web to learn human activities for interactive systems [Fast *et al.*, 2016b]. Augur's evaluation indicated that with regard to low-level behaviors such as actions, these data provide a surprisingly accurate mirror of human behavior. Empath contributes a different perspective, that text on the web can be an appropriate tool for learning a breadth of topical and emotional categories, to the benefit of social science. In other research communities, systems have used unsupervised models to capture emergent practice in open source code [Fast *et al.*, 2014] or design [Kumar *et al.*, 2013]. In Empath, we adapt these techniques to mine natural language for its relation to emotional and topical categories.

Finally, Empath also benefits from prior work in commonsense knowledge representation. Existing databases of linguistic and commonsense knowledge provide networks of facts that computers should know about the world [Liu and Singh, 2004; Miller, 1995; Esuli and Sebastiani, 2006]. We draw on some of this knowledge, like the ConceptNet hierarchy, when seeding Empath's categories. Further, Empath itself captures a set of relations on the topical and emotional connotations of words. Some aspects of these connotations may be mineable from social media, if they are of the sort that people are likely to advertise on Twitter [Kiciman, 2015].

3 Empath Applications

Here we motivate the value of Empath through two example analyses that illustrate its breadth and flexibility.

3.1 Understanding deception in hotel reviews

What kinds of words accompany our lies? For our first example, we used Empath to analyze a dataset of deceptive hotel reviews reported previously by Ott el al. [Ott *et al.*, 2011]. This dataset contains 3200 truthful hotel reviews from TripAdvisor and deceptive reviews created by workers on Amazon Mechanical Turk. The original study found that liars tend to write more imaginatively and use less concrete language.

Exploring the deception dataset

We ran Empath's full set of categories over the truthful and deceptive reviews, and produced aggregate statistics for each. Using normalized means of the category counts for each group, we then computed odds ratios and p-values for the categories most likely to appear in deceptive and truthful reviews. All the results we report are significant after a Bonferroni correction ($\alpha = 2.5e^{-5}$).

Our results provide new evidence in support of the Ott et al. study, suggesting that deceptive reviews convey stronger sentiment across both positively and negatively charged categories, and tend towards exaggerated language. The liars more often use language that is *tormented* (2.5 odds) or *joyous* (2.3 odds), for example "it was **torture** hearing the sounds of the elevator which just would never stop" or "I got a **great** deal and I am so **happy** that I stayed here." The truth-tellers more often discuss concrete ideas and phenomena like the *ocean* (1.6 odds,), *vehicles* (1.7 odds) or *noises* (1.7 odds), for example "It seemed like a nice enough place with reasonably close **beach** access" or "they took forever to Valet our **car**." We see a tendency towards more mundane activities among the truth-tellers through categories like *eating* (1.3 odds), *cleaning* (1.3 odds), or *hygiene* (1.2 odds). "I ran the **shower** for ten minutes without ever receiving any hot water." For the liars interactions seem to be more evocative, involving *death* (1.6 odds) or *partying* (1.3 odds). "The **party** that keeps you awake will not be your favorite band practicing for their next **concert**."

For exploratory research questions, Empath provides a high-level view over many potential categories, some of which a researcher may not have thought to investigate. Lying hotel reviewers, for example, may not have realized they give themselves away by fixating on *smell* (1.4 odds), "the room was **pungent** with what **smelled** like human excrement", or their overuse of emotional terms, producing significantly higher odds ratios for 13 of Empath's 32 emotional categories. Truthful reviews, on the other hand, display higher odds ratios for none of Empath's emotional categories.

Spatial language in lies

While the original study provided some evidence that liars use less spatially descriptive language, it wasn't able to test the theory directly. Using Empath, we can generate a new set of human validated terms that capture this idea, creating a new *spatial* category. To do so, we tell Empath to seed the category with the terms "big", "small", and "circular". Empath then discovers a series of related terms and uses the crowd to validate them, producing the cluster:

> circular, small, big, large, huge, gigantic, tiny, rectangular, rectangle, massive, giant, enormous, smallish, rounded, middle, oval, sized, size, miniature, circle, colossal, center, triangular, shape, boxy, ...

When we then add the new *spatial* category to our analysis, we find it favors truthful reviews by 1.2 odds ($p < 0.001$). Truth-tellers use more spatial language, for example, "the room that we originally were in had a **huge square** cut out of the wall that had exposed pipes, bricks, dirt and dust." In aggregate, liars are not as apt in these concrete details.

3.2 Mood on twitter and time of day

For our second example, we used Empath to replicate the relationship between mood on twitter and time of day demonstrated by Golder and Macy [Golder and Macy, 2011]. The corpus of tweets analyzed by the original paper is not publicly available, so we reproduced its findings on a smaller corpus of 591,520 tweets, running LIWC as a benchmark (Figure 2).

The original paper shows a low of negative sentiment in the morning that rises over the rest of the day. We find a similar relationship on our data with both Empath and LIWC: a low in the morning (around 8am), peaking to a high around 11pm. The signals reported by Empath and LIWC over each hour are strongly correlated (r=0.90). Using a 1-way ANOVA to test for changes in mean negative affect by hour, Empath reports a highly significant difference ($F(23, 591520) = 17.2$, $p < 0.001$), as does LIWC ($F = 6.8$, $p < 0.001$). For positive sentiment, Empath and LIWC again replicate similarly with strong correlation between tools (r=0.87). Both tools once more report highly significant ANOVAs by hour: Empath $F = 5.9, p < 0.001$; LIWC $F = 7.3, p < 0.001$.

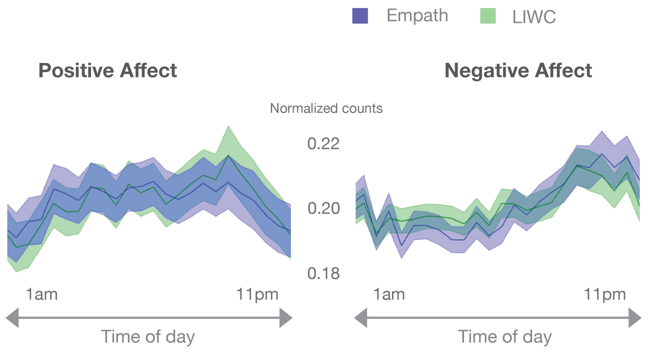

Figure 2: Empath replicates the work of Golder and Macy, investigating how mood on Twitter relates to time of day. Empath and LIWC are strongly correlated over hours for positive (r=0.87) and negative (r=0.90) sentiment.

4 Empath

Empath analyzes text across hundreds of topics and emotions. Like LIWC and other dictionary-based tools, it counts category terms in a text document. However, Empath covers a broader set of categories than other tools, and users can generate and validate new categories with a few seed words.

4.1 Designing Empath's categories

Empath provides 200 human validated categories, which cover topics like *violence*, *depression*, or *femininity*. We drew these categories from common concepts in the ConceptNet knowledge base and Parrott's hierarchy of emotions [Shaver *et al.*, 1987]. While Empath's topical and emotional categories stem from different sources of knowledge, we generate member terms for both kinds of categories in the same way. Given a set of seed terms (from ConceptNet or the Parrott hierarchy), Empath learns from a large corpus of text to predict and validate hundreds of similar categorical terms.

We generate category terms by querying a vector space model trained by a neural network on a large corpus of text. This model allows Empath to examine the similarity between words across many dimensions of meaning. For example, given seed words like "facebook" and "twitter,' Empath finds related terms like "pinterest" and "selfie."

Training a neural word embedding model

To train Empath's model, we adapt the skip-gram architecture introduced by Mikolov et al. [Mikolov *et al.*, 2013a]. This is an unsupervised learning that teaches a neural network to predict co-occurring words in a corpus. For example, the network might learn that "death" predicts a nearby occurrence of the word "carrion," but not of "incest." Over training the network learns a representation of each word that is predictive of its context, and we can then borrow these representations, called neural embeddings, to map words onto a vector space.

More formally, for word w and context C in a network with negative sampling, a skip-gram network will learn weights that maximize the dot product $w \cdot w_c$ and minimize $w \cdot w_n$ for $w_c \in C$ and w_n sampled randomly from the vocabulary. The context C of a word is determined by a sliding window over the document, of a size typically in (0,7).

We train our network on data from Wattpad, Reddit, and the New York Times [Fast *et al.*, 2016b; Fast and Horvitz, 2016a; 2016b]. The network uses a hidden layer of 150 neurons (which defines the dimensionality of the embedding space), a sliding window size of five, a minimum word count of thirty (i.e., a word must occur at least thirty times to appear in the training set), negative sampling, and down-sampling of frequent terms. These techniques reflect current best practices in language modeling [Mikolov *et al.*, 2013b].

Building categories with a vector space

We use the neural embeddings created by our skip-gram network to construct a vector space model (VSM). Similar models trained on neural embeddings, such as word2vec, enable powerful forms of analogous reasoning (e.g., the vector arithmetic for the terms "King - Man + Queen" produces a vector close to "Woman") [Luo and Xu, 2015]. In our case, VSMs allow Empath to discover member terms for categories.

VSMs encode concepts as vectors, where each dimension of the vector $v \in \mathbb{R}^n$ conveys a feature relevant to the concept. For Empath, each vector v is a word, and each of its dimensions defines the weight of its connection to one of the hidden layer neurons (the neural embeddings). The space is $\mathbb{M}(n \times h)$ where n is the size of our vocabulary (40,000), and h the number of hidden nodes in the network (150).

Empath's VSM selects member terms for its categories (e.g., social media, violence, shame) by using cosine similarity, a similarity measure over vector spaces, to find nearby terms in the space. Concretely, we search the vector spaces on multiple seed terms by querying on the vector sum of those terms—a kind of reasoning by analogy. From a small seed of words, Empath can gather hundreds of terms related to a given category, and then use these terms for textual analysis.

4.2 Refining categories with crowd validation

Human-validated categories can ensure that accidental terms do not slip into a lexicon. By filtering Empath's categories through the crowd, we offer the benefits of both modern NLP and human validation: increasing category precision, and more carefully validating category contents.

To validate each of Empath's categories, we created a crowdsourcing pipeline on Amazon Mechanical Turk [Fast *et al.*, 2016a]. We divided the total number of words to be filtered across many separate tasks, where each task consists of twenty words to be rated for a given category. For each of these words, workers select a relationship on a four point scale: not related, weakly related, related, and strongly related. We ask three independent workers to complete each task at a cost of \$0.14 per task. Prior work has shown that three workers are enough for reliable results in labeling tasks, given high quality contributors [Sheng *et al.*, 2008]. So, if we want to filter a category of 200 words, we would have $200/20 = 10$ tasks, which must be completed by three workers, at a total cost of $10 * 3 * 0.14 = \$4.2$ for this category. We limit tasks to Masters workers to ensure quality and aggregate crowdworker feedback by majority vote. Workers demonstrated high agreement on the labeling task (81%).

5 Comparison of Empath and LIWC

How well do categories generated by Empath's unsupervised model approximate gold standard lexicons created by humans? To find out, we selected 12 categories from LIWC and compared their category word counts with Empath over a large, mixed-corpus dataset.

First, we created a mixed text dataset evenly divided among tweets [Mohammad *et al.*, 2014], StackExchange opinions [Danescu-Niculescu-Mizil *et al.*, 2013], movie reviews [Pang *et al.*, 2002], hotel reviews [Ott *et al.*, 2011], and chapters sampled from four classic novels on Project Gutenberg (David Copperfield, Moby Dick, Anna Karenina, and The Count of Monte Cristo) [Gutenberg, 2016]. This mixed corpus contains more than 2 million words in total across 4500 individual documents.

For each LIWC category under analysis, we chose up to 5 seed words that allowed Empath to best approximate the category. We selected these seed words through manual discovery on a training sample of the corpus and applied crowd validation to these categories. On the held out test dataset, we then collected category counts for both LIWC and Empath and compared the resulting Pearson correlation coefficient (PCC) values, with and without crowd validation.

We find that Empath's categories are highly correlated with LIWC's, with average PCCs of 0.906 (with crowd validation) and 0.90 (without crowd validation). The lowest correlation is 0.86 (against LIWC's *work* category), and the highest correlation is 0.944 (against *positive emotion*). A more thorough analysis of these relationships is available in the expanded version of this paper [Fast *et al.*, 2016a].

6 Conclusion

Empath aims to combine modern NLP techniques with the transparency of dictionaries like LIWC. In doing so, it provides both broader and deeper forms of text analysis than existing tools. In breadth, Empath offers hundreds of predefined lenses through which researchers can analyze text. In depth, its user-defined categories provide a flexible means by which researchers can ask domain-specific questions. These questions are ever changing, as is our use of language. Empath is a living lexicon—able to keep up with each.

References

[Bradley and Lang, 1999] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. In *Technical Report C-1, The Center for Research in Psychophysiology, University of Florida*, 1999.

[Chambers and Jurafsky, 2009] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In *ACL*, 2009.

[Danescu-Niculescu-Mizil *et al.*, 2013] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *ACL*, 2013.

[Esuli and Sebastiani, 2006] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, 2006.

[Fast and Horvitz, 2016a] Ethan Fast and Eric Horvitz. Identifying dogmatism in social media: Signals and models. *EMNLP*, 2016.

[Fast and Horvitz, 2016b] Ethan Fast and Eric Horvitz. Long-term trends in the public perception of artificial intelligence. *AAAI*, 2016.

[Fast *et al.*, 2014] Ethan Fast, Daniel Steffe, Lucy Wang, Michael Bernstein, and Joel Brandt. Emergent, crowd-scale programming practice in the ide. In *CHI*, 2014.

[Fast *et al.*, 2016a] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *CHI*, pages 4647–4657. ACM, 2016.

[Fast *et al.*, 2016b] Ethan Fast, Will McGrath, Pranav Rajpurkar, and Michael Bernstein. Mining human behaviors from fiction to power interactive systems. In *CHI*, 2016.

[Golder and Macy, 2011] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. In *Science*, volume 333, pages 1878–1881, 2011.

[Gutenberg, 2016] Gutenberg. Project gutenberg. In *https://www.gutenberg.org/*, 2016.

[Hatzivassiloglou and McKeown, 1997] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[Hutto and Gilbert, 2014] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI*, 2014.

[Kiciman, 2015] Emre Kiciman. Towards learning a knowledge base of actions from experiential microblogs. In *AAAI Spring Symposium*, 2015.

[Kramer *et al.*, 2014] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. In *Proceedings of the National Academy of Sciences*, volume 111, pages 8788–8790, 2014.

[Kumar *et al.*, 2013] Ranjitha Kumar, Arvind Satyanarayan, Cesar Torres, Maxine Lim, Salman Ahmad, Scott R Klemmer, and Jerry O Talton. Webzeitgeist: Design Mining the Web. In *CHI*, 2013.

[Liu and Singh, 2004] H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit. In *BT Technology Journal*, 2004.

[Luo and Xu, 2015] Qun Luo and Weiran Xu. Learning word vectors efficiently using shared representations and document representations. In *AAAI*, 2015.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, 2013.

[Miller, 1995] George A. Miller. WordNet: A lexical database for english. In *In Commun. ACM*, 1995.

[Mohammad and Turney, 2013] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence*, volume 29, pages 436–465, 2013.

[Mohammad *et al.*, 2014] Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. In *Information Processing & Management*. Elsevier, 2014.

[Ott *et al.*, 2011] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*, 2011.

[Pang *et al.*, 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*, 2002.

[Pennebaker *et al.*, 2001] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count. In *Mahway: Lawrence Erlbaum Associates*, 2001.

[Shaver *et al.*, 1987] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. In *Journal of personality and social psychology*, volume 52, page 1061. American Psychological Association, 1987.

[Sheng *et al.*, 2008] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *SIGKDD*, 2008.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[Stone *et al.*, 1966] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. MIT press, 1966.

[Turney, 2002] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *ACL*, 2002.