Combining Sentiment Lexica with a Multi-View Variational Autoencoder

Alexander Hoyle\textsuperscript{3} Lawrence Wolf-Sonkin\textsuperscript{1} Hanna Wallach\textsuperscript{5} Ryan Cotterell\textsuperscript{6} Isabelle Augenstein\textsuperscript{7}
\textsuperscript{3}University College London, London, UK
\textsuperscript{1}Department of Computer Science, Johns Hopkins University, Baltimore, USA
\textsuperscript{5}Microsoft Research, New York City, USA
\textsuperscript{6}Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
\textsuperscript{7}Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
alexander.hoyle.17@ucl.ac.uk, lawrencews@jhu.edu
hanna@dirichlet.net, rdc42@cam.ac.uk, augenstein@di.ku.dk

Abstract

When assigning quantitative labels to a dataset, different methodologies may rely on different scales. In particular, when assigning polarities to words in a sentiment lexicon, annotators may use binary, categorical, or continuous labels. Naturally, it is of interest to unify these labels from disparate scales to both achieve maximal coverage over words and to create a single, more robust sentiment lexicon while retaining scale coherence. We introduce a generative model of sentiment lexica to combine disparate scales into a common latent representation. We realize this model with a novel multi-view variational autoencoder (VAE), called SentiVAE. We evaluate our approach via a downstream text classification task involving nine English-Language sentiment analysis datasets; our representation outperforms six individual sentiment lexica, as well as a straightforward combination thereof.

1 Introduction

Sentiment lexica provide an easy way to automatically label texts with polarity values, and are also frequently transformed into features for supervised models, including neural networks (Palogiannidi et al., 2016; Ma et al., 2018). Indeed, given their utility, a veritable cottage industry has emerged focusing on the design of sentiment lexica. In practice, using any single lexicon, unless specifically and carefully designed for the particular domain of interest, has several downsides. For example, any lexicon will typically have low coverage compared to the language’s entire vocabulary, and may have misspecified labels for the domain. In many cases, it may therefore be desirable to combine multiple sentiment lexica into a single representation. Indeed, some research on unifying such lexica has emerged (Emerson and Declerck, 2014; Altrabsheh et al., 2017), borrowing ideas from crowdsourcing (Raykar et al., 2010; Hovy et al., 2013). However, this is a non-trivial task, because lexica can use binary, categorical, or continuous scales to quantify polarity—in addition to different interpretations for each—and thus cannot easily be combined. In Fig. 1, we show an example of the same word labeled using different lexica to illustrate the nature of the challenge.

To combine sentiment lexica with disparate scales, we introduce SentiVAE, a novel multi-view variant of the variational autoencoder (VAE) (Kingma and Welling, 2014). SentiVAE, visualized as a graphical model in Fig. 2, differs from the original VAE in two ways: (i) it uses a Dirichlet latent variable (rather than a Gaussian) for each word in the combined vocabulary, and (ii) it has multiple emission distributions—one for each lexicon. Because the latent variables are shared across the lex-
We use the following commonly used English-language sentiment lexica: SentiWordNet (Baccianella et al., 2010), MPQA (Wilson et al., 2005), SenticNet 5 (Cambria et al., 2014), Hu-Liu (Hu and Liu, 2004), GI (Stone et al., 1962), and VADER (Hutto and Gilbert, 2014). Descriptive statistics for each lexicon are shown in Tab. 1. Each word in SentiWordNet is labeled with two real values, each in the interval $[-1, 1]$, corresponding to the strength of positive and negative sentiment (e.g., the label $(0, 0)$ is neutral, while the label $(1, 0)$ is maximally positive). Each word in VADER is labeled by ten different human evaluators, with each evaluator providing a polarity value on a nine-point scale (where the midpoint is neutral), yielding a 10-dimensional label. MPQA, Hu-Liu, and GI all use binary scales. Lastly, each word in SenticNet is labeled with a real value in the interval $[-1, 1]$, where 0 is neutral.

### 3 SentiVAE

We first describe a figurative generative process for a single sentiment lexicon $d \in \mathcal{D}$, where $\mathcal{D}$ is a set of sentiment lexica. Imagine there is a true (latent) polarity value $z^w$ associated with each word $w$ in the lexicon’s vocabulary. When the lexicon’s creator labels that word according to their chosen scale (e.g., thumbs-up or thumbs-down, a real value in the interval $[0, 1]$), they deterministically transform this true value to their chosen scale via a function $f(\cdot; \theta_d)$. Sometimes, noise is introduced during this labeling process, corrupting the label as it leaves the ethereal realm and producing the (observed) polarity label $x^w_d$. They then add this potentially noisy label to the lexicon.

Given a lexicon of observed polarity labels, the latent polarity values can be inferred using a VAE. The original VAE posits a generative model of observed data $\mathcal{X}$ and latent variables $\mathcal{Z}$: $P(\mathcal{X}, \mathcal{Z}) = P(\mathcal{X} | \mathcal{Z}) P(\mathcal{Z})$. Inference of $\mathcal{Z}$ then proceeds by approximating the (intractable) posterior $P(\mathcal{Z} | \mathcal{X})$ with a Gaussian distribution, factorized over the individual latent variables. A parameterized encoder function compresses $\mathcal{X}$ into $\mathcal{Z}$, while a parameterized decoder function reconstructs $\mathcal{X}$ from $\mathcal{Z}$.

SentiVAE extends the original VAE model to combine multiple lexica with disparate scales, producing a common latent representation of the polarity value for each word in the combined vocabulary.

#### Generative process

Given a set of sentiment lexica $\mathcal{D}$ with a combined vocabulary $\mathcal{W}$, SentiVAE posits a common latent representation $z^w$ of the polarity value for each word $w \in \mathcal{W}$, where $z^w$ is a three-dimensional categorical distribution over...
the sentiments positive, negative, and neutral.

The generative process starts by drawing each latent polarity value \( z^w \) from a three-dimensional Dirichlet prior, parameterized by \( \alpha^w = (1, 1, 1) \):

\[
z^w \sim \text{Dir}(\alpha^w).
\]

If the word is uncontroversial, we spurt this prior somewhat using the number of lexica in which the word appears \( c(w) \). Specifically, we add \( c(w) \) to the parameter for the sentiment associated with that word in the lexica, e.g., \( \alpha^{\text{SUPERB}} = (1 + c(\text{SUPERB}), 1, 1) \). This has the effect of regularizing the inferred latent polarity value toward the desired distribution over sentiments.

Having generated \( z^w \), the process proceeds by “decoding” \( z^w \) into each lexicon’s chosen scale. First, for each lexicon \( d \in D \), \( z^w \) is deterministically transformed via neural network \( f(\cdot; \theta_d) \) with a single 32-dimensional hidden layer, parameterized by lexicon-specific weights \( \theta_d \):

\[
\rho_d^w = f(z^w; \theta_d).
\]

The transformed value \( \rho_d^w \) is then used to generate the (observed) polarity label \( x_d^w \) for that lexicon:

\[
x_d^w \sim P_d(\alpha_d^w \mid \rho_d^w).
\]

The dimensionality of \( \rho_d^w \) and the emission distribution \( P_d \) are lexicon-specific. For SentiWordNet, \( P_d \)

is a two-dimensional Gaussian with mean \( \mu_d^w \) and a diagonal covariance matrix equal to 0.01 \( I \); for VADER, \( P_d \) consists of ten nine-dimensional categorical distributions, collectively parameterized by \( \rho_d^w \); for MPQA, Hu-Liu, and GI, \( P_d \) is a Bernoulli distribution, parameterized by \( \mu_d^w \); and for SentiNet, \( P_d \) is a univariate Gaussian with mean and variance each an element in a two-dimensional \( \rho_d^w \).

**Inference.** Inference involves forming the posterior distribution over the latent polarity values \( Z \) given the observed polarity labels \( X \). Because computing the normalizing constant \( P(X) \) is intractable, we instead approximate the posterior with a family of distributions \( Q_{\lambda}(Z) \), indexed by variational parameters \( \lambda \). Specifically, we use

\[
Q_{\lambda}(Z) = \prod_{w \in W} Q_w(z^w) = \prod_{w \in W} \text{Dir}(\beta^w). 
\]

To construct \( \beta^w \), we first define a neural network \( g(\cdot; \phi_d) \), with a single 32-dimensional hidden layer, which “encodes” \( x_d^w \) into a three-dimensional vector. The output of this neural network is then transformed via a softmax as follows:

\[
\omega_d^w = \text{softmax}(g(x_d^w; \phi_d)) \quad (5)
\]

\[
\beta^w = 1 + \sum_{d \in D} \omega_d^w. \quad (6)
\]

The intuition behind \( \beta_w \) can be understood by appealing to the “pseudocount” interpretation of Dirichlet parameters. Each lexicon contributes exactly one pseudocount, divided among positive, negative, and neutral, to what would otherwise be a symmetric, uniform Dirichlet distribution. As a consequence of this construction, words that appear in more lexica will have more concentrated Dirichlets. Intuitively, this property is appealing.

We optimize the resulting ELBO objective (Blei et al., 2017) with respect to the variational parameters via stochastic variational inference (Hoffman

### Table 2: Descriptive statistics for the training portions of the sentiment analysis datasets. \( N \): number of texts.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>( N )</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>Movies</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>Yelp</td>
<td>Product reviews</td>
<td>100000</td>
<td>5 / 3</td>
</tr>
<tr>
<td>SemEval</td>
<td>Twitter</td>
<td>7668</td>
<td>3</td>
</tr>
<tr>
<td>MultiDom</td>
<td>Product reviews</td>
<td>6500</td>
<td>2</td>
</tr>
<tr>
<td>ACL</td>
<td>Scientific</td>
<td>248</td>
<td>5 / 3</td>
</tr>
<tr>
<td>ICLR</td>
<td>Scientific</td>
<td>2166</td>
<td>10 / 3</td>
</tr>
</tbody>
</table>
Table 3: Classification accuracies for our representation, six lexica, and a straightforward combination thereof.

<table>
<thead>
<tr>
<th></th>
<th>IMDB 2C</th>
<th>Yelp 5C</th>
<th>Yelp 3C</th>
<th>SemEval 3C</th>
<th>MultiDom 2C</th>
<th>ACL 5C</th>
<th>ACL 3C</th>
<th>ICLR 10C</th>
<th>ICLR 3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiVAE $E_q(z^w)$</td>
<td>72.7</td>
<td>49.8</td>
<td>57.5</td>
<td>46.0</td>
<td>70.8</td>
<td>66.7</td>
<td>73.3</td>
<td>92.6</td>
<td>87.0</td>
</tr>
<tr>
<td>SentiVAE $\beta^w$</td>
<td>73.4</td>
<td>49.7</td>
<td>59.4</td>
<td>52.2</td>
<td>74.7</td>
<td>73.3</td>
<td>80.0</td>
<td>92.6</td>
<td>86.5</td>
</tr>
<tr>
<td>SentiWordNet</td>
<td>63.4</td>
<td>36.0</td>
<td>47.6</td>
<td>32.2</td>
<td>62.0</td>
<td>60.0</td>
<td>53.3</td>
<td>89.1</td>
<td>83.5</td>
</tr>
<tr>
<td>MPQA</td>
<td>65.4</td>
<td>44.0</td>
<td>53.0</td>
<td>29.9</td>
<td>67.4</td>
<td>60.0</td>
<td>53.3</td>
<td>89.1</td>
<td>83.5</td>
</tr>
<tr>
<td>SenticNet</td>
<td>60.5</td>
<td>38.4</td>
<td>43.4</td>
<td>37.2</td>
<td>62.3</td>
<td>60.0</td>
<td>53.3</td>
<td>89.1</td>
<td>83.9</td>
</tr>
<tr>
<td>Hu-Liu</td>
<td>67.2</td>
<td>46.6</td>
<td>56.4</td>
<td>31.5</td>
<td>69.4</td>
<td>60.0</td>
<td>53.3</td>
<td>89.1</td>
<td>83.5</td>
</tr>
<tr>
<td>GI</td>
<td>58.4</td>
<td>40.7</td>
<td>47.9</td>
<td>31.3</td>
<td>61.6</td>
<td>60.0</td>
<td>53.3</td>
<td>89.1</td>
<td>83.5</td>
</tr>
<tr>
<td>VADER</td>
<td>71.7</td>
<td>46.8</td>
<td>59.3</td>
<td>38.5</td>
<td>73.5</td>
<td>66.7</td>
<td>66.7</td>
<td>94.3</td>
<td>86.1</td>
</tr>
<tr>
<td>Combined</td>
<td>75.6</td>
<td>51.0</td>
<td>64.1</td>
<td>50.6</td>
<td>75.4</td>
<td>66.7</td>
<td>66.7</td>
<td>93.9</td>
<td>86.1</td>
</tr>
</tbody>
</table>

4 Experiments and Results

To evaluate our approach, we first use SentiVAE to combine the six lexica described in §2. For each word $w$ in the combined vocabulary, we obtain an estimate of $z^w$ by taking the mean of $Q_{\beta^w}(z^w) = \text{Dir}(\beta^w)$—i.e., by normalizing $\beta^w$. We compare this representation to using $\beta^w$ directly, because $\beta^w$ contains information about SentiVAE’s certainty about the word’s latent polarity value. We evaluate our common latent representation via a text classification task involving nine English-language sentiment analysis datasets: IMDB (Maas et al., 2011), Yelp (Zhang et al., 2015), SemEval 2017 Task 4 (SemEval, Rosenthal et al. (2017)), multi-domain sentiment analysis (MultiDom, Blitzer et al. (2007)), and PeerRead (Kang et al., 2018) with splits ACL 2017 and ICLR 2017 (Kang et al., 2018). Each dataset consists of multiple texts (e.g., tweets, articles), each labeled with an overall sentiment (e.g., positive). Descriptive statistics for each dataset are shown in Tab. 2.

For each dataset, we transform each text into an average polarity value using either our representation, one of the six lexica, or a straightforward combination thereof, where the polarity value for each word in the (combined) vocabulary is a 16-dimensional vector that consists of a concatenation of polarity values. (Unlike SentiVAE, this concatenation does not yield a single sentiment lexicon that retains scale coherence, while achieving maximal coverage over words.) Specifically, we replace each token with its corresponding polarity value, and then average the these values (Go et al., 2009; Özdemir and Bergler, 2015; Kiritchenko et al., 2014). We then use the training portion of the dataset to learn a logistic regression classifier to predict the overall sentiment of each text from its average polarity value. Finally, we use the testing portion to compute the accuracy of the classifier.

Results. The results in Tab. 3 show that our representation using $\beta^w$ outperforms the individual lexica for all but one dataset, and that our representation using the mean of $Q_{\beta^w}(z^w)$ outperforms them for six datasets. This is likely because SentiVAE has a richer representation of sentiment than any individual lexicon, and it has greater coverage over words (see Tab. 4). The results in Tab. 5 support the former reason: even when we limit the words in our representation to match those in an individual lexicon, our representation still outperforms the individual lexicon. Unsurprisingly, our representation especially outperforms lexica with unidimensional scales. We also find that our representation outperforms the straightforward combination for datasets from domains that are not well supported by the individual lexica (see Tabs. 1 and 2 for lexicon and dataset sources, respectively). By combining lexica from different domains, our representation captures a general notion of sentiment that is not tailored to any specific domain.

5 Conclusion

We introduced a generative model of sentiment lexica to combine disparate scales into a common
Table 4: Coverage over words (percentage) by lexicon for the training portions of four of the nine datasets.

<table>
<thead>
<tr>
<th></th>
<th>IMDB</th>
<th>SemEval</th>
<th>Multi</th>
<th>ICLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiVAE</td>
<td>70</td>
<td>64</td>
<td>81</td>
<td>71</td>
</tr>
<tr>
<td>SentiWordNet</td>
<td>15</td>
<td>14</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>MPQA</td>
<td>10</td>
<td>7</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>SenticNet</td>
<td>40</td>
<td>39</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>Hu-Liu</td>
<td>7</td>
<td>5</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>GI</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>VADER</td>
<td>7</td>
<td>6</td>
<td>13</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5: Classification accuracies for a 10% validation portion of two of the datasets. The first row, labeled SentiVAE, contains the classification accuracy for our representation using $\beta_w$. Subsequent (lexicon-specific) rows compare our representation (SV), restricted to the vocabulary of that lexicon, to the lexicon itself (Lex).

<table>
<thead>
<tr>
<th></th>
<th>IMDB 2C</th>
<th>SemEval 3C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SV Lex</td>
<td>SV Lex</td>
</tr>
<tr>
<td>SentiVAE</td>
<td>74.7</td>
<td>–</td>
</tr>
<tr>
<td>SentiWordNet</td>
<td>70.6</td>
<td>63.4</td>
</tr>
<tr>
<td>MPQA</td>
<td>73.5</td>
<td>66.6</td>
</tr>
<tr>
<td>SenticNet</td>
<td>74.4</td>
<td>60.9</td>
</tr>
<tr>
<td>Hu-Liu</td>
<td>73.6</td>
<td>68.4</td>
</tr>
<tr>
<td>GI</td>
<td>71.4</td>
<td>59.3</td>
</tr>
<tr>
<td>VADER</td>
<td>73.6</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Acknowledgements

We would like to thank to Adam Forbes for the design of Fig. 1. We further acknowledge the support of the NVIDIA Corporation with the donation of the Titan Xp GPU used to conduct this research.

References


ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 168–177. ACM.

C. J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14).


