What Affects Patient (Dis)satisfaction?
Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model

Michael J. Paul
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
mpaul@cs.jhu.edu

Byron C. Wallace
Center for Evidence-based Medicine
Brown University
Providence, RI 02903
byron_wallace@brown.edu

Mark Dredze
Human Language Technology
Center of Excellence
Johns Hopkins University
Baltimore, MD 21211
mdredze@cs.jhu.edu

Abstract
We analyze patient reviews of doctors using a novel probabilistic joint model of topic and sentiment based on factorial LDA (Paul and Dredze 2012). We leverage this model to exploit a small set of previously annotated reviews to automatically analyze the topics and sentiment latent in over 50,000 online reviews of physicians (and we make this dataset publicly available). The proposed model outperforms baseline models for this task with respect to model perplexity and sentiment classification. We report the most representative words with respect to positive and negative sentiment along three clinical aspects, thus complementing existing qualitative work exploring patient reviews of physicians.

Introduction and motivation
Individuals are increasingly turning to the web for healthcare information. A recent survey (Fox and Duggan 2013) found that 72% of internet users have looked online for health information in the past year, and one in five for reviews of particular treatments or doctors. In a random sample of 500 urologists, online reviews were found to have been written about ~80% of them (Ellimoottil et al. 2012). These numbers will likely increase in coming years.

The shift toward online health information consumption and sharing has produced a proliferation of health-related user-generated content, including online doctor reviews. Such reviews have clear value to patients, but they are also valuable in that taken en masse they may reveal insights into factors that affect patient satisfaction. In an analysis of online healthcare provider reviews, López et al. (2012) noted that comments regarding interpersonal manner and technical competence tended to be more positive, whereas comments about systems issues (e.g., regarding office staff) tended to be more mixed. Elsewhere, Segal et al. (2012) have shown online doctor reviews can track quality of care.

A drawback to existing explorations of online provider reviews (with the exception of (Brody and Elhadad 2010)) is that they have been qualitative in nature. This approach limits the potential scope of analysis, and has precluded conduct of the sort of larger-scale analyses necessary to comprehensively elucidate the content of online doctor reviews.

Further, qualitative analyses rely on subjective judgements and summarizations of reviews, whereas we prefer a formal model that facilitates discovery of patterns from data.

To this end, we propose a joint probabilistic model that captures both the sentiment and aspects latent in the free text of online provider reviews. Our aim is to elucidate the factors that most affect consumer sentiment regarding interactions with their doctor. We have compiled a dataset of over 50,000 online doctor reviews, which we make publicly available. Reviews include ratings along various aspects: staffing, helpfulness and knowledgeability. Our aim is to tease out the text associated with strong sentiment along these aspects to illustrate the factors that most influence patient satisfaction. We develop a novel probabilistic generative model based on Factorial Latent Dirichlet Allocation (f-LDA) (Paul and Dredze 2012) that exploits both the review ratings and a small set of manual annotations to uncover salient language representative of negative and positive sentiment along different aspects of provider care (e.g., interpersonal manner, technical competence, etc.).

In contrast to previous work in this direction (Brody and Elhadad 2010), our model jointly captures both aspect and sentiment in healthcare provider reviews. Further, rather than taking a fully unsupervised approach, we leverage a small set of manual annotations created for a previously conducted qualitative study of online provider reviews (López et al. 2012) to bias the model parameter estimates. To validate this approach we show that the proposed model can be used to predict aspect sentiment ratings on held-out reviews with greater accuracy than models without such priors.
Technical competence

<table>
<thead>
<tr>
<th>positive</th>
<th>negative</th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>shows empathy, professional, communicates well</td>
<td>poor listener, judgmental, racist</td>
<td>good decision maker, follows up on issues, knowledgeable</td>
<td>poor decision maker, prescribes the wrong medication, disorganized</td>
</tr>
<tr>
<td>friendly staff, short wait times, convenient location</td>
<td>difficult to park, rude staff, expensive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Illustrative tags underneath the three main aspects identified in (López et al. 2012).

Related work

We begin by reviewing two disparate threads of related work: (1) explorations of online doctor reviews, and, (2) models of text that jointly account for aspect and sentiment.

Online doctor reviews There has been a flurry of recent research concerning online physician-rating websites (Segal et al. 2012; Emmert, Sander, and Pisch 2013; López et al. 2012; Galizzi et al. 2012; Ellimoottil et al. 2012). We have already discussed the work by López et al. (2012): perhaps their most interesting finding was that reviews often concern aspects beyond the patient-doctor relationship (e.g., office staff). They also found that well-perceived bedside manner was a key to successful patient-doctor interactions.

In more quantitative work, Segal et al. (2012) analyzed the relationship between high-volume surgeons (who perform many operations) and online reviews. Noting that surgeons who perform more procedures tend to have better clinical outcomes and safety records, they found that they could both identify high-volume surgeons using online reviews, and that high-volume surgeons tended to receive more praise.

Similar to our application, Brody and Elhadad (2010) explored “salient aspects” in online reviews of healthcare providers via Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). They partitioned reviews according to provider specialty (e.g., ob-gyn, dentist) and discovered aspects with LDA run (with modifications) on each of the specialties. In contrast to our work, however, they did not exploit existing labels or aspects; their aspects are discovered topics. Moreover, whereas they used a variant of LDA, we explicitly model the topic and sentiment jointly to discover salient factors with respect to negative and positive reviews across different aspects of treatment.

Joint models of topic and sentiment There has been some work on the task of aspect-based sentiment summarization (Mei et al. 2007; Titov and McDonald 2008), a variation of the general task of sentiment classification (Pang and Lee 2004), in which the aim is to classify documents (e.g., movie reviews) as containing ‘positive’ or ‘negative’ sentiment. Generally, this is with respect to the overall sentiment in a document whereas aspect-sentiment models focus latent sentiment on specific aspects. For example, a restaurant review may praise the food but lament the service (positive for the ‘food’ aspect but negative for ‘service’).

Most work in this direction relies on the machinery of LDA, a probabilistic generative model of text (Blei, Ng, and Jordan 2003). Extending LDA to account for aspect-specific sentiment, Titov and McDonald (2008) considered the general task of jointly modeling text and aspect ratings for sentiment summarization, exploiting supervision by leveraging existing aspect labels. Mei et al. (2007) proposed a mixture model that combines topic and sentiment components through a switch variable; Lu et al. (2009) used a similar approach to summarize sentiment of eBay feedback.

An advantage of the factorial model we use is that we jointly model the interaction of topic and sentiment together; that is, we model that some words are associated with the particular pairing of a topic category and sentiment polarity value. As we will show in the next section, our model’s use of rich priors is also easily extendable to incorporating prior knowledge, such as labeled data from domain experts and review metadata.

A Joint Topic-Sentiment Model

We propose a novel topic-sentiment model based on factorial LDA (Paul and Dredze 2012). Our approach jointly models topic and sentiment, allowing us to analyze the positive and negative words associated with specific topics such as interpersonal manners or technical competence. We first review factorial LDA and then describe two extensions to tailor the model to our doctor review analysis.

First, we use the approach of Paul and Dredze (2013) to leverage annotations provided by content experts, thus exploiting these for larger-scale inference. Second, we introduce a novel extension to the model that incorporates tangential but informative user ratings to guide the model.

Background: Factorial LDA Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a generative model...
of text in which words in a document reflect a mixture of latent topics. Each token is associated with a latent “topic” variable $z$. Factorial LDA (f-LDA) (Paul and Dredze 2012) generalizes LDA to allow each token to be associated with a $K$-dimensional vector of latent variables $\zeta$, rather than a single topic. We consider a two-dimensional model in which each token is associated with two variables corresponding to topic and sentiment. As we describe f-LDA, we will focus on this version with two dimensions, called factors.

In LDA, each document has a distribution $\theta^{(d)}$ over topics (we will use ‘topic’ and ‘aspect’ interchangeably), while under our two-dimensional f-LDA model, each document has a distribution $\theta^{(d)}$ over all possible (topic, sentiment) pairs. In LDA, each topic $z$ is associated with a distribution over words $\phi_z$; our f-LDA model has a word distribution $\phi_\zeta$ for every (topic, sentiment) pair. In this model’s generative story, words are generated by first sampling a pair $\zeta = (t_1, t_2)$ from the document’s pair distribution $\theta^{(d)}$, then sampling a word from that pair’s word distribution $\phi_\zeta$.

So far there is nothing that models the notion of two separate dimensions. Intuitively, one would expect commonalities across the various (topic, sentiment) pairs that share a topic or sentiment value. For example, the pairs (INTERPERSONAL,POSITIVE) and (INTERPERSONAL,NEGATIVE) should both have words pertaining to a doctor’s interpersonal manners, even though they are associated with two different word distributions. Similarly, even though the pairs (INTERPERSONAL,POSITIVE) and (SYSTEMS,POSITIVE) are about two different aspects, they both represent positive sentiment, and so they should both contain positive words.

The key ingredient of f-LDA is that the Dirichlet priors for $\theta$ and $\phi$ share parameters for pairs that share components. For example, all pairs with POSITIVE sentiment include positive-specific parameters in the prior for those pairs’ word distributions $\phi_\zeta$. Formally, $\phi_{\zeta}$ (the word distribution for pair $\zeta$) has a Dirichlet($\phi^{(\zeta)}$) prior, where for each word $w$, $\omega_w^{(\zeta)}$ is a log-linear function:

$$\omega_w^{(\zeta)} \propto \exp\left(\omega^{(B)} + \omega^{(0)} + \omega_{1w}^{(topic)} + \omega_{2w}^{(sentiment)}\right)$$  
(1)

where $\omega^{(B)}$ is a corpus-wide precision scalar (the bias), $\omega^{(0)}$ is a corpus-specific bias for word $w$, and $\omega_{kw}^{(k)}$ is a bias parameter for word $w$ for component $t_k$ of the $k$th factor. That is, each topic and sentiment has a weight vector over the vocabulary, and the prior for a particular pair is influenced by the weight vectors of each of the two factors. Thus, all Dirichlet parameters for pairs with a particular topic or sentiment will be influenced by that topic or sentiment’s weight vector, encouraging commonalities across pairs. The standard version of f-LDA assumes the $\omega$ parameters are all independent and normally distributed around 0.

The prior over each document’s distribution over pairs has a similar log-linear prior, where weights for each factor are combined to influence the distribution. $\theta^{(d)}$ is drawn from Dirichlet($\alpha^{(d)}$), with $\alpha_{\zeta}$ for each pair $\zeta$ defined as:

$$\alpha_{\zeta}^{(d)} = b_\zeta \exp\left(\alpha^{(B)} + \alpha_{1\zeta}^{(D.top.)} + \alpha_{1\zeta}^{(d.top.)} + \alpha_{2\zeta}^{(D.sent.)} + \alpha_{2\zeta}^{(d.sent.)}\right)$$  
(2)

Similar to before, $\alpha^{(B)}$ is a global bias parameter, while the $\alpha^{(d)}$ vectors are corpus-wide weights and $\alpha^{(d)}$ are document-specific weights. Structuring the prior in this way models the intuition that, for example, if the positive sentiment is prevalent in a document, it is a priori likely across all topics. Finally, $b_\zeta$ is a real-valued scalar in $(0, 1)$ which acts as a sparsity pattern over the space of pairs: the intuition is that certain (topic, sentiment) combinations may have very low probability across the entire corpus, so the model can learn near-zero $b$ values for such pairs. In our doctor review experiments, the $b$ values are always close to 1, so we will not dwell on these sparsity variables.

Posterior inference and parameter estimation utilize a collapsed Gibbs sampler, which samples values of the latent $\zeta$ variables, and gradient ascent algorithm, in which the various $\alpha$ and $\omega$ hyperparameters are optimized. See Paul and Dredze (2012) for more details.

**Priors from Labeled Data**

Recall that one of our datasets (López et al. 2012) contains annotations for both the aspects of interest and sentiment. Because the number of labeled reviews is relatively small (less than 1K compared to over 50K), we still want to leverage the unannotated data, rather than simply using a fully supervised model. Paul and Dredze (2013) showed how to incorporate labeled data into f-LDA in a semi-supervised manner, using the labeled data to create priors.

The idea is to train a similar but simplified model on the labeled data, and then use the supervised parameters as priors over the f-LDA parameters. In particular, we use a supervised variant of SAGE (Eisenstein, Ahmed, and Xing 2011), which gives the following model of our data:

$$P(\text{word} \, w | \text{topic} = i, \text{sentiment} = j) = \frac{\exp(\eta_{iw}^{(\text{background})} + \eta_{iw}^{(\text{topic})} + \eta_{iw}^{(\text{sentiment})})}{\sum_{w'} \exp(\eta_{iw'}^{(\text{background})} + \eta_{iw'}^{(\text{topic})} + \eta_{iw'}^{(\text{sentiment})})}$$  
(3)

This log-linear model has a similar form as Equation 1, but the topic and sentiment labels are fixed across the entire document, rather than being unique to each token, and it represents a probability rather than a Dirichlet vector. We fix the background $\eta$ vectors to be the observed vector of corpus log-frequencies over the vocabulary, which acts as an “overall” weight vector, and we estimate the $\eta$ weight vectors for each topic and sentiment from the labeled data. These parameters are then used as the means of the Gaussian priors over $\omega$ for the background and each factor $k$, i.e.:

$$\omega_{iw}^{(0)} \sim N(\eta_{iw}^{(\text{background})}, \sigma^2); \omega_{iw}^{(k)} \sim N(\eta_{iw}^{(k)}, \sigma^2)$$

The $\eta$ parameters are learned using gradient ascent.

We use the three high-level topic labels described above from the López et al. dataset: interpersonal manner (INTERPERSONAL), technical competence (TECHNICAL), and systems issues (SYSTEMS) (Table 2). Each review in the labeled
Priors from User Ratings

The reviews in the RateMDs corpus contain user ratings (integers ranging from 1 to 5) for three categories: knowledgeability, staff, helpfulness.\(^2\) As a novel extension to f-LDA for the purpose of this topic-sentiment task, we attempt to leverage these ratings (which are not quite what we want to model, but provide valuable side information) to further guide the model in inferring the different topic and sentiment pairs. In this section, we show how to incorporate these user ratings into the document priors.

These rating categories naturally correspond to similar labels as in the López et al. dataset, albeit only roughly. We created the following category-to-topic mapping:

- ‘Knowledgeability’: TECHNICAL
- ‘Staff’: SYSTEMS
- ‘Helpfulness’: INTERPERSONAL

For each pair \( t \) in document \( d \), we use the user ratings to create rating variables \( r_{t}^{(d)} \) centered around the middle value of 3: for each topic, we set the value of \( r_{t}^{(d)} \) for the positive sentiment to be the original user rating minus 3, while the \( r_{t}^{(d)} \) value for the negative sentiment is the negation of the positive. For example, if the user rating for ‘Staff’ was 2, then \( r_{\text{SYSTEMS,POS}}^{(d)} = -1 \) and \( r_{\text{SYSTEMS,NEG}}^{(d)} = 1 \), while if the user rating for ‘Helpful’ was 3, then the \( r_{t}^{(d)} \) variables for both the positive and negative INTERPERSONAL pairs would be 0. These \( r \) variables can thus be used to bias the document’s pair distribution toward or away from pairs that have a high or low user rating. The \( r \) values are simply set to 0 for topics beyond the first 3, for which we do not have ratings.

We incorporate \( r_{t}^{(d)} \) into the document’s prior over pair distributions, so that topics with high ratings are \textit{a priori} more likely to contain that topic paired with positive sentiment and less likely to contain that topic paired with negative sentiment. Specifically, we modify the log-linear equation in Eq. 2 to include an additional term containing \( r_{t}^{(d)} \).

\[
\exp\left(\alpha_{t}^{(\text{B})} + \alpha_{t}^{(\text{D,top})} + \alpha_{t}^{(d,top)} + \alpha_{t}^{(D,\text{sen.})} + \alpha_{t}^{(d,\text{sen.})} + \rho r_{t}^{(d)}\right)
\]  

(4)

where \( \rho > 0 \) is a scaling parameter that controls how strongly the rating variable should influence the prior.

We optimize \( \rho \) to maximize likelihood. For mathematical convenience, we first re-parameterize \( \rho \) as \( \exp(\tilde{\rho}) \), allowing us to optimize \( \tilde{\rho} \in \mathbb{R} \) rather than \( \rho \in (0, \infty) \). We also place a regularization prior on \( \tilde{\rho} \): a 0-mean Gaussian. The partial derivative of the corpus likelihood with respect to \( \tilde{\rho} \) is:

\[
\frac{\partial \ell}{\partial \tilde{\rho}} = -\frac{\tilde{\rho}}{\sigma^2} + \sum_{d} \sum_{t} r_{t}^{(d)} \exp(\tilde{\rho}) \alpha_{t}^{(d)} \times \left(\Psi(n_{t}^{d} + \tilde{\alpha}_{t}^{(d)}) - \Psi(\tilde{\alpha}_{t}^{(d)}) + \Psi(\sum_{z} \alpha_{z}^{(d)}) - \Psi(\sum_{z} n_{z}^{d} + \tilde{\alpha}_{z}^{(d)})\right)
\]  

(5)

where \( n_{t}^{d} \) is the number of times the pair \( t \) appeared in document \( d \), given the current state of the Gibbs sampler. We optimize this with gradient ascent along with the other hyperparameters of the model.

Experiments and Analysis

Our model utilizes two extensions to f-LDA. In our experiments, we compare this full model to ablated versions:

- ‘B’: baseline model without extensions;
- ‘W’: model with word priors from labeled data;
- ‘R’: model with document priors from user ratings;
- ‘WR’: full model with both extensions.

We also compared against LDA with comparable numbers of word distributions. For example, when comparing against f-LDA with 3 topics, we use 6 topics in LDA, because f-LDA in this case has 6 word distributions for the 3 topics paired with 2 sentiment values. All of our Gibbs samplers are run for 5000 iterations with a gradient ascent step size of \( 10^{-3} \). The variance of the Gaussian prior over the parameters was \( \sigma^2 = 1 \) for \( \alpha \) and \( \rho \), \( \sigma^2 = 0.5 \) for \( \omega \). LDA was run for the same number of iterations; the Dirichlet hyperparameters were optimized for likelihood.

We initialized \( \alpha^{(B)} = -2 \) and \( \omega^{(B)} = -6 \), the other \( \omega \) parameters were initialized to their corresponding \( \eta \) values when applicable, and all other hyperparameters were initialized to 0. Finally, to tilt the model parameters slightly toward the correct sentiment values, we initialized \( \alpha_{\text{POS}}^{(d,\text{sen.})} = 0.1 \) if the average user rating across the three categories was \( \geq 3 \) and \(-0.1 \) otherwise, with \( \alpha_{\text{NEG}}^{(d,\text{sen.})} = -\alpha_{\text{POS}}^{(d,\text{sen.})} \).

When training SAGE on the labeled data, our gradient ascent algorithm was run for 1000 iterations with a step size of \( 10^{-2} \). The Gaussian prior variance over the \( \eta \) was \( \sigma^2 = 0.1 \).

Evaluation

Results are from 5 fold cross-validation where within each fold, we perform 10 inference trials through randomly initialized sampling chains on the training set (80% of the data) and selecting inferred parameters with the lowest perplexity on the held-out set (20% of the data). For inference on the held-out set, we fix all except for document-specific parameters. We run the sampler for 1000 iterations, and then average the parameters sampled from 100 iterations. No information about the user ratings is used during inference on the test set; the ‘R’ extensions used by the models only apply during training.
Rating prediction  To evaluate learning we predict the user ratings for the three categories described above. If the model discovers meaningful topics, w.r.t. aspects and sentiment, then encoding reviews as distributions ought to improve sentiment predictions along these related aspects.

To verify that this is indeed the case, we encoded each review by its distribution over the inferred (topic, sentiment) pairs. Specifically, we represent each review as a vector of length $2 \times |Z|$, representing every (topic, sentiment) pair. We set entry $j$ for each review equal to the proportion of words in the review assigned to pair $j$. Because we have an ordinal outcome (ratings are integers from 1 (negative) to 5 (positive)), we used ordinal logistic regression to predict ratings. We also experimented with using standard Bag-of-Words feature encoding, and mixtures of BoW and the topic distribution representation.

The proposed full model almost always outperforms the other models (Figure 1), and the models with extensions almost always outperform the baseline f-LDA model in terms of prediction. All f-LDA models can predict the user ratings substantially better than LDA. Additionally, the two ‘W’ models typically had lower variance than others, perhaps because the word priors lead to more consistency in the inferred parameters. Exact numbers for prediction from the topic output are shown in Table 3.

Model analysis

Our model uncovers several interesting salient patterns (Figure 2). Consider the general negative and positive terms:

We also evaluated the models by examining the perplexity of held-out data, but the results were very inconsistent, and no model variant was significantly better than others.

<table>
<thead>
<tr>
<th></th>
<th>staff</th>
<th>helpfulness</th>
<th>knowledgeability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.27</td>
<td>1.52</td>
<td>1.36</td>
</tr>
<tr>
<td>LDA</td>
<td>1.50</td>
<td>1.60</td>
<td>1.43</td>
</tr>
<tr>
<td>B</td>
<td>0.95</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>W</td>
<td>0.91</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>R</td>
<td>0.99</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td>WR</td>
<td>0.91</td>
<td>0.59</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Table 3: Mean absolute error of rating prediction using topics as features with $Z=3$.

rude and asked are the top two most negative tokens, highlighting the importance of communicative/interpersonal skills. Indeed, it would seem that poor communicative skills is the most oft generally complained about aspect of patient care. Generally positive terms are (unsurprisingly) dominated by superlatives (e.g., wonderful). Additionally, we find that the words associated with topics generally match what one would expect: the interpersonal topic includes words like manner and caring; the technical topic contains words about surgeries and other operations; and the systems topic contains words about the hospital and office, such as appointment, staff, and nurse.

Increasing beyond 3 topics yield more specific words in each topic. The interpersonal topic with $Z=9$ included the words unprofessional, arrogant, attitude, cold, and condescending (negative), along with compassionate and understanding (positive). When examining the topics beyond the first 3, we find that the model learns clusters of more specific topics such as dentistry and family matters, but some of these topics are noisier and the positive and negative distributions for the same topic index sometimes do not even correspond to the same theme. The fact that the topics beyond the first 3 are less salient may explain why the rating prediction with f-LDA was generally worse for $Z>3$.
Certain issues appear to be associated with specific polar sentiment. For example, *medication* and *prescription* is mentioned more in negative contexts – patients remark when they get a wrong prescription, but a correct prescription is unremarkable. Bedside manners are primarily mentioned in positive contexts. Systems issues related to appointments and wait times are primarily mentioned in negative contexts; this agrees with López et al.’s remark that many patients were concerned with wait times (2012).

We observed sentiment-specific differences in the language used by patients to reference their doctor: the word *dr* has a high prior for the positive sentiment, and upon inspection we noticed that users writing positive reviews were more likely to mention the doctor by name and title (“Dr. X”), while addressing the doctor by name only (“X”) or no name (“s/he,” “the doctor”) was more common with negative reviews. More generally, we noticed that specific mentions of people appear in positive contexts. For example, the technical/operations topic includes many words describing family members (*husband, daughter*). In the systems issues topic, *staff* has a higher probability in the positive distribution than *office* (a more abstract institution), whereas this pattern is reversed in the negative distribution.

### Conclusions and Future Directions

We have analyzed a large collection of online physician reviews with a modified version of factorial LDA (Paul and Dredze 2012). We enriched the model by incorporating a small amount of labeled training data (using ideas from previous work) and by incorporating aspect-specific user ratings from the review metadata (a novel extension created for this project). Our experimental results have demonstrated the quality and predictiveness of this new model. Quantitatively, we showed that our model is much more predictive of aspect ratings than alternative models, and qualitatively we verified that the model is learning sensible (topic, sentiment) pairs. Now that we have demonstrated that our model learns the information we care about in this corpus, we intend to use it for more experiments in the future. For example, we can use it for extractive summarization (Paul and Dredze 2013), providing context to enable deeper analysis.

### References


