Abstract

We introduce a new dataset for multi-class emotion analysis from long-form narratives in English. The Dataset for Emotions of Narrative Sequences (DENS) was collected from both classic literature available on Project Gutenberg and modern online narratives available on Wattpad, annotated using Amazon Mechanical Turk. A number of statistics and baseline benchmarks are provided for the dataset. Of the tested techniques, we find that the fine-tuning of a pre-trained BERT model achieves the best results, with an average micro-F1 score of 60.4%. Our results show that the dataset provides a novel opportunity in emotion analysis that requires moving beyond existing sentence-level techniques.

1 Introduction

Humans experience a variety of complex emotions in daily life. These emotions are heavily reflected in our language, in both spoken and written forms.

Many recent advances in natural language processing on emotions have focused on product reviews (McAuley et al., 2015) and tweets (Mohammad et al., 2018). These datasets are often limited in length (e.g. by the number of words in tweets), purpose (e.g. product reviews), or emotional spectrum (e.g. binary classification).

Character dialogues and narratives in storytelling usually carry strong emotions. A memorable story is often one in which the emotional journey of the characters resonates with the reader. Indeed, emotion is one of the most important aspects of narratives. In order to characterize narrative emotions properly, we must move beyond binary constraints (e.g. good or bad, happy or sad).

In this paper, we introduce the Dataset for Emotions of Narrative Sequences (DENS) for emotion analysis, consisting of passages from long-form fictional narratives from both classic literature and modern stories in English. The data samples consist of self-contained passages that span several sentences and a variety of subjects. Each sample is annotated by using one of 9 classes and an indicator for annotator agreement.

2 Background

Using the categorical basic emotion model (Plutchik, 1979), (Mohammad and Kiritchenko, 2015; Mohammad, 2012) studied creating lexicons from tweets for use in emotion analysis. Recently, (Mohammad et al., 2018), (Klinger et al., 2018) and (Kant et al., 2018) proposed shared-tasks for multi-class emotion analysis based on tweets.

Fewer works have been reported on understanding emotions in narratives. Emotional Arc (Reagan et al., 2016) is one recent advance in this direction. The work used lexicons and unsupervised learning methods based on unlabelled passages from titles in Project Gutenberg.

For labelled datasets on narratives, (Alm et al., 2005) provided a sentence-level annotated corpus of children’s stories and (Kim and Klinger, 2018) provided phrase-level annotations on selected Project Gutenberg titles.

To the best of our knowledge, the dataset in this work is the first to provide multi-class emotion labels on passages, selected from both Project Gutenberg and modern narratives. The dataset is available upon request for non-commercial, research only purposes.

3 Dataset

In this section, we describe the process used to collect and annotate the dataset.

---

1https://www.gutenberg.org/

2Please send requests to: academic_dataset@wattpad.com
3.1 Plutchik’s Wheel of Emotions

The dataset is annotated based on a modified Plutchik’s wheel of emotions.

The original Plutchik’s wheel consists of 8 primary emotions: Joy, Sadness, Anger, Fear, Anticipation, Surprise, Trust, Disgust. In addition, more complex emotions can be formed by combing two basic emotions. For example, Love is defined as a combination of Joy and Trust (Fig. 1).

![Figure 1: Plutchik’s wheel of emotions](Wikimedia, 2011)

The intensity of an emotion is also captured in Plutchik’s wheel. For example, the primary emotion of Anger can vary between Annoyance (mild) and Rage (intense).

We conducted an initial survey based on 100 stories with a significant fraction sampled from the romance genre. We asked readers to identify the major emotion exhibited in each story from a choice of the original 8 primary emotions.

We found that readers have significant difficulty in identifying Trust as an emotion associated with romantic stories. Hence, we modified our annotation scheme by removing Trust and adding Love. We also added the Neutral category to denote passages that do not exhibit any emotional content.

The final annotation categories for the dataset are: Joy, Sadness, Anger, Fear, Anticipation, Surprise, Love, Disgust, Neutral.

3.2 Passage Selection

We selected both classic and modern narratives in English for this dataset. The modern narratives were sampled based on popularity from Wattpad. We parsed selected narratives into passages, where a passage is considered to be eligible for annotation if it contained between 40 and 200 tokens.

In long-form narratives, many non-conversational passages are intended for transition or scene introduction, and may not carry any emotion. We divided the eligible passages into two parts, and one part was pruned using selected emotion-rich but ambiguous lexicons such as cry, punch, kiss, etc. Then we mixed this pruned part with the unpruned part for annotation in order to reduce the number of neutral passages. See Appendix A.1 for the lexicons used.

3.3 Mechanical Turk (MTurk)

MTurk was set up using the standard sentiment template and instructed the crowd annotators to ‘pick the best/major emotion embodied in the passage’.

We further provided instructions to clarify the intensity of an emotion, such as: “Rage/Annoyance is a form of Anger”, “Serenity/Ecstasy is a form of Joy”, and “Love includes Romantic/Family/Friendship”, along with sample passages.

We required all annotators have a ‘master’ MTurk qualification. Each passage was labelled by 3 unique annotators. Only passages with a majority agreement between annotators were accepted as valid. This is equivalent to a Fleiss’s $\kappa$ score of greater than 0.4.

For passages without majority agreement between annotators, we consolidated their labels using in-house data annotators who are experts in narrative content. A passage is accepted as valid if the in-house annotator’s label matched any one of the MTurk annotators’ labels. The remaining passages are discarded. We provide the fraction of annotator agreement for each label in the dataset.

During the annotation process, several annotators had suggested for us to include additional emotions such as confused, pain, and jealousy,
### Genre Distribution (%)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Distribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mystery/Thriller</td>
<td>19.7</td>
</tr>
<tr>
<td>Paranormal</td>
<td>16.6</td>
</tr>
<tr>
<td>Fantasy</td>
<td>13.2</td>
</tr>
<tr>
<td>Horror</td>
<td>11.3</td>
</tr>
<tr>
<td>Romance</td>
<td>8.7</td>
</tr>
<tr>
<td>Action/Adventure</td>
<td>5.3</td>
</tr>
<tr>
<td>Other</td>
<td>9.3</td>
</tr>
</tbody>
</table>

Table 1: Genre distribution of the modern narratives which are common to narratives. As they were not part of the original Plutchik’s wheel, we decided not to include them. An interesting future direction is to study the relationship between emotions such as ‘pain versus sadness’ or ‘confused versus surprise’ and improve the emotion model for narratives.

### 3.4 Dataset Statistics

The dataset contains a total of 9,710 passages, with an average of 6.24 sentences per passage, 16.16 words per sentence, and an average length of 86 words.

The vocabulary size is 28K (when lowercased). It contains over 1,600 unique titles across multiple categories, including 88 titles (1,520 passages) from Project Gutenberg. All of the modern narratives were written after the year 2000, with notable amount of themes in *coming-of-age*, *strong-female-lead*, and *LGBTQ+.* The genre distribution is listed in Table 1.

In the final dataset, 21.0% of the data has consensus between all annotators, 73.5% has majority agreement, and 5.48% has labels assigned after consultation with in-house annotators.

The distribution of data points over labels with top lexicons (lower-cased, normalized) is shown in Table 2. Note that the *Disgust* category is very small and should be discarded. Furthermore, we suspect that the data labelled as *Surprise* may be noisier than other categories and should be discarded as well.

Table 3 shows a few examples labelled data from classic titles. More examples can be found in Table 6 in the Appendix A.2.

### 4 Benchmarks

We performed benchmark experiments on the dataset using several different algorithms. In all experiments, we have discarded the data labelled with *Surprise* and *Disgust*.

We pre-processed the data by using the SpaCy\(^3\) pipeline. We masked out named entities with entity-type specific placeholders to reduce the chance of benchmark models utilizing named entities as a basis for classification.

Benchmark results are shown in Table 4. The dataset is approximately balanced after discarding the *Surprise* and *Disgust* classes. We report the average micro-F1 scores, with 5-fold cross validation for each technique.

We provide a brief overview of each benchmark experiment below. Among all of the benchmarks, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) achieved the best performance with a 0.604 micro-F1 score.

Overall, we observed that deep-learning based techniques performed better than lexical based methods. This suggests that a method which attends to context and themes could do well on the dataset.

#### 4.1 Bag-of-Words-based Benchmarks

We computed bag-of-words-based benchmarks using the following methods:

- Classification with TF-IDF + Linear SVM (TF-IDF + SVM)
- Classification with Depeche++ Emotion lexicons (Araque et al., 2018) + Linear SVM (Depeche + SVM)
- Classification with NRC Emotion lexicons (Mohammad and Turney, 2010, 2013) + Linear SVM (NRC + SVM)
- Combination of TF-IDF and NRC Emotion lexicons (TF-NRC + SVM)

#### 4.2 Doc2Vec + SVM

We also used simple classification models with learned embeddings. We trained a Doc2Vec model (Le and Mikolov, 2014) using the dataset and used the embedding document vectors as features for a linear SVM classifier.

#### 4.3 Hierarchical RNN

For this benchmark, we considered a Hierarchical RNN, following (Sordoni et al., 2015). We used two BiLSTMs (Graves et al., 2005) with 256 units.

\(^3\)https://spacy.io/
I found this was a little too close upon him, but I made it up in what follows. He stood stock-still for a while and said nothing, and I went on thus: “You cannot,” says I, “without the highest injustice, believe that I yielded upon all these persuasions without a love not to be questioned, not to be shaken again by anything that could happen afterward. If you have such dishonourable thoughts of me, I must ask you what foundation in any of my behaviour have I given for such a suggestion?”

She stretched hers eagerly and gratefully towards him. What had happened? Through all the numbness of her blood, there sprang a strange new warmth from his strong palm, and a pulse, which she had almost forgotten as a dream of the past, began to beat through her frame. She turned around all a-tremble, and saw his face in the glow of the coming day.

Ah! That moving procession that has left me by the road-side! Its fantastic colors are more brilliant and beautiful than the sun on the undulating waters. What matter if souls and bodies are failing beneath the feet of the ever-pressing multitude! It moves with the majestic rhythm of the spheres. Its discordant clashes sweep upward in one harmonious tone that blends with the music of other worlds—to complete God’s orchestra.

<table>
<thead>
<tr>
<th>Label</th>
<th>Gutenberg</th>
<th>Total</th>
<th>Top Lexicons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>318</td>
<td>1711</td>
<td>take, love, long, really, want, always, though, away, look</td>
</tr>
<tr>
<td>Fear</td>
<td>159</td>
<td>1412</td>
<td>left, behind, right, want, let, death, go, say, think</td>
</tr>
<tr>
<td>Sadness</td>
<td>195</td>
<td>1402</td>
<td>father, always, little, look, something, us, really, mother, think</td>
</tr>
<tr>
<td>Anger</td>
<td>192</td>
<td>1306</td>
<td>feel, much, well, man, look, us, say, something, love</td>
</tr>
<tr>
<td>Joy</td>
<td>241</td>
<td>1266</td>
<td>see, always, let, long, make, hand, away, get, really</td>
</tr>
<tr>
<td>Love</td>
<td>162</td>
<td>1157</td>
<td>hand, know, right, let, happy, get, ever, us, look</td>
</tr>
<tr>
<td>Anticipation</td>
<td>147</td>
<td>1020</td>
<td>know, long, life, make, get, think, blood, want, feel</td>
</tr>
<tr>
<td>Surprise</td>
<td>102</td>
<td>362</td>
<td>love, find, looking, know, well, much, something, door, really</td>
</tr>
<tr>
<td>Disgust</td>
<td>4</td>
<td>74</td>
<td>get, hand, inside, let, hate, table, men, always, make</td>
</tr>
</tbody>
</table>

Table 2: Dataset label distribution

<table>
<thead>
<tr>
<th>Text</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>I found this was a little too close upon him, but I made it up in what follows. He stood stock-still for a while and said nothing, and I went on thus: “You cannot,” says I, “without the highest injustice, believe that I yielded upon all these persuasions without a love not to be questioned, not to be shaken again by anything that could happen afterward. If you have such dishonourable thoughts of me, I must ask you what foundation in any of my behaviour have I given for such a suggestion?”</td>
<td>Angry</td>
</tr>
<tr>
<td>She stretched hers eagerly and gratefully towards him. What had happened? Through all the numbness of her blood, there sprang a strange new warmth from his strong palm, and a pulse, which she had almost forgotten as a dream of the past, began to beat through her frame. She turned around all a-tremble, and saw his face in the glow of the coming day.</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Ah! That moving procession that has left me by the road-side! Its fantastic colors are more brilliant and beautiful than the sun on the undulating waters. What matter if souls and bodies are failing beneath the feet of the ever-pressing multitude! It moves with the majestic rhythm of the spheres. Its discordant clashes sweep upward in one harmonious tone that blends with the music of other worlds—to complete God’s orchestra.</td>
<td>Joy</td>
</tr>
</tbody>
</table>

Table 3: Sample data from classic titles

<table>
<thead>
<tr>
<th>Model</th>
<th>micro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF + SVM</td>
<td>0.450</td>
</tr>
<tr>
<td>Depeche + SVM</td>
<td>0.254</td>
</tr>
<tr>
<td>NRC + SVM</td>
<td>0.286</td>
</tr>
<tr>
<td>TF-NRC + SVM</td>
<td>0.458</td>
</tr>
<tr>
<td>Doc2Vec + SVM</td>
<td>0.403</td>
</tr>
<tr>
<td>HRNN</td>
<td>0.469</td>
</tr>
<tr>
<td>BiRNN + Self-Attention</td>
<td>0.487</td>
</tr>
<tr>
<td>ELMo + BiRNN</td>
<td>0.516</td>
</tr>
<tr>
<td>Fine-tuned BERT</td>
<td><strong>0.604</strong></td>
</tr>
</tbody>
</table>

Table 4: Benchmark results (averaged 5-fold cross-validation)

4.4 Bi-directional RNN and Self-Attention (BiRNN + Self-Attention)

One challenge with RNN-based solutions for text classification is finding the best way to combine word-level representations into higher-level representations.
Self-attention (Yang et al., 2016; Lin et al., 2017; Sinha et al., 2018) has been adapted to text classification, providing improved interpretability and performance. We used (Lin et al., 2017) as the basis of this benchmark.

The benchmark used a layered Bi-directional RNN (60 units) with GRU cells and a dense layer. Both self-attention layers were 60 units in size and cross-entropy was used as the cost function.

Note that we have omitted the orthogonal regularizer term, since this dataset is relatively small compared to the traditional datasets used for training such a model. We did not observe any significant performance gain while using the regularizer term in our experiments.

4.5 ELMo embedding and Bi-directional RNN (ELMo + BiRNN)

Deep Contextualized Word Representations (ELMo) (Peters et al., 2018) have shown recent success in a number of NLP tasks. The unsupervised nature of the language model allows it to utilize a large amount of available unlabelled data in order to learn better representations of words.

We used the pre-trained ELMo model (v2) available on Tensorhub4 for this benchmark. We fed the word embeddings of ELMo as input into a one layer Bi-directional RNN (16 units) with GRU cells (with dropout) and a dense layer. Cross-entropy was used as the cost function.

4.6 Fine-tuned BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) has achieved state-of-the-art results on several NLP tasks, including sentence classification.

We used the fine-tuning procedure outlined in the original work to adapt the pre-trained uncased BERTLARGE5 to a multi-class passage classification task. This technique achieved the best result among our benchmarks, with an average micro-F1 score of 60.4%.

5 Conclusion

We introduce DENS, a dataset for multi-class emotion analysis from long-form narratives in English. We provide a number of benchmark results based on models ranging from bag-of-word models to methods based on pre-trained language models (ELMo and BERT).

Our benchmark results demonstrate that this dataset provides a novel challenge in emotion analysis. The results also demonstrate that attention-based models could significantly improve performance on classification tasks such as emotion analysis.

Interesting future directions for this work include: 1. incorporating common-sense knowledge into emotion analysis to capture semantic context and 2. using few-shot learning to bootstrap and improve performance of underrepresented emotions.

Finally, as narrative passages often involve interactions between multiple emotions, one avenue for future datasets could be to focus on the multi-emotion complexities of human language and their contextual interactions.

References


Subjectivity, Sentiment and Social Media Analysis, pages 31–42. Brussels, Belgium. Association for Computational Linguistics.


Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. Computational Intelligence, 31(2):301–326.


