DENS: A Dataset for Multi-class Emotion Analysis

Chen Liu and **Muhammad Osama** and **Anderson de Andrade**
Wattpad
Toronto, ON, Canada
`cecilia, muhammad.osama, anderson@wattpad.com`

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; Kant 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[1].

For labelled datasets on narratives, (Alm et al., 2005) provided a sentence-level annotated corpus of childrens' 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[2].

3 Dataset

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

---

[1] https://www.gutenberg.org/
[2] Please send requests to: academic_dataset@wattpad.com

3.1 Plutchiks Wheel of Emotions

The dataset is annotated based on a modified Plutchiks wheel of emotions.

The original Plutchiks 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).
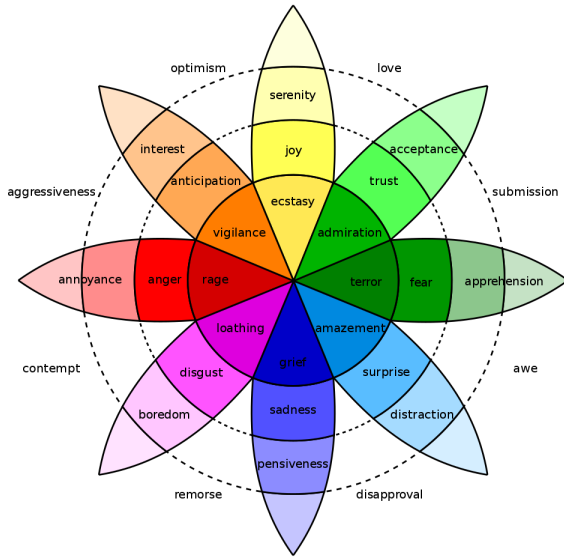

Figure 1: Plutchik's wheel of emotions (Wikimedia, 2011)

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 κ 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.

Though passages may lose some emotional context when read independently of the complete narrative, we believe annotator agreement on our dataset supports the assertion that small excerpts can still convey coherent emotions.

During the annotation process, several annotators had suggested for us to include additional emotions such as confused, pain, and jealousy,

Genre	Distribution (%)
Mystery/Thriller	19.7
Paranormal	16.6
Fantasy	13.2
Horror	11.3
Romance	8.7
Action/Adventure	5.3
Other	9.3

Table 1: Genre distribution of the modern narratives

which are common to narratives. As they were not part of the original Plutchiks wheel, we decided to not 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 9710 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 1600 unique titles across multiple categories, including 88 titles (1520 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 each to model sentences and documents. The tokens of a sentence were processed independently

---

[3] https://spacy.io/

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

Table 2: Dataset label distribution

Text	Label
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?"	Angry
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.	Anticipation
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.	Joy

Table 3: Sample data from classic titles

Model	micro-F1
TF-IDF + SVM	0.450
Depeche + SVM	0.254
NRC + SVM	0.286
TF-NRC + SVM	0.458
Doc2Vec + SVM	0.403
HRNN	0.469
BiRNN + Self-Attention	0.487
ELMo + BiRNN	0.516
Fine-tuned BERT	**0.604**

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

of other sentence tokens. For each direction in the token-level BiLSTM, the last outputs were concatenated and fed into the sentence-level BiLSTM as inputs.

The outputs of the BiLSTM were connected to 2 dense layers with 256 ReLU units and a Softmax layer. We initialized tokens with publicly available embeddings trained with GloVe (Pennington et al., 2014). Sentence boundaries were provided by SpaCy. Dropout was applied to the dense hidden layers during training.

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 *Tensorhub*[4] 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 BERT$_{\text{LARGE}}$[5] 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.

---

[4]https://tfhub.dev/google/elmo/2
[5]https://tfhub.dev/google/bert_uncased_L-24_H-1024_A-16/1

References

Cecilia Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2018. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359. Association for Computational Linguistics.

Roman Klinger, Orphee De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding.

Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Saif M. Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)*. 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.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Robert Plutchik. 1979. *Emotions: A general psychoevolutionary theory*, volume 1. Psycology Press,Taylor and Francis Group.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5:31.

Koustuv Sinha, Yue Dong, Jackie Chi Kit Cheung, and Derek Ruths. 2018. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM.

Wikimedia. 2011. Robert plutchik's wheel of emotions. File: Plutchik-wheel.svg.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

A Appendices

A.1 Lexicons

cry	punch	blood	knife
flower	moon	wind	exclaim
chuckle	tear	punch	yell
kiss	touch	warm	dead
shiver	chill		

Table 5: Lexicons used to prune part of the data for labelling

A.2 Sample Data

Table 6 shows sample passages from classic titles with corresponding labels.

Text	Label
He took his screwdriver and again took off the lid of the coffin. Arthur looked on, very pale but silent. When the lid was removed he stepped forward. He evidently did not know that there was a leaden coffin, or at any rate, had not thought of it. When he saw the rent in the lead, the blood rushed to his face for an instant, but as quickly fell away again, so that he remained of a ghastly whiteness. He was still silent. Van Helsing forced back the leaden flange, and we all looked in and recoiled.	Fear
The chair went to matchwood at the bottom, and we rolled apart into the gutter. He sprang to his feet, waving his fists and wheezing like an asthmatic. "Had enough?" he panted. "You infernal bully!" I cried, as I gathered myself together.	Anger
The judges sat grave and mute, gave me an easy hearing, and time to say all that I would, but, saying neither Yes nor No to it, pronounced the sentence of death upon me, a sentence that was to me like death itself, which, after it was read, confounded me. I had no more spirit left in me, I had no tongue to speak, or eyes to look up either to God or man.	Sadness
The Prince burst into a yelling, shrieking fit of laughter. Instantly the yellow-haired serfs in waiting, the Calmucks at the hall-door, and the half-witted dwarf who crawled around the table in his tow shirt, began laughing in chorus, as violently as they could. The Princess Martha and Prince Boris laughed also; and while the old man's eyes were dimmed with streaming tears of mirth, quickly exchanged nods. The sound extended all over the castle, and was heard outside of the walls.	Joy
"Do not be such an unreasonable child", he remonstrated, feebly. "I do not love you with the wild, irrational passion of former years; but I have the tenderest regard for you, and my heart warms at the sight of your sweet face, and I shall do all in my power to make you as happy as any man can make you who–"	Love
I looked around for his birds, and not seeing them, asked him where they were. He replied, without turning round, that they had all flown away. There were a few feathers about the room and on his pillow a drop of blood. I said nothing, but went and told the keeper to report to me if there were anything odd about him during the day.	Neutral

Table 6: Sample data from classic titles