Sentiment Analysis of Turtle Books: A 439/539 NLP Project

Jane Smith
University of Arizona, Tucson, AZ, USA.
jane@arizona.edu

Abstract

The abstract should be approximately 100 words, and outline the problem (1-sentence), the specific task you're working on (1-sentence), your proposed approach/solution to the task (1 to 2 sentences), and your results (1-sentence). Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

1 Introduction

The introduction should describe the problem area that you're working on in high-level general terms and it's utility, identify that gap in knowledge/the need, and then how you address that need.

For example: Sentiment analysis aims to automatically identify whether a piece of text is describing generally positive, neutral, or negative emotion *(the high-level description)*. Sentiment analysis has broad utility, for example in automatically determining whether natural language comments left by consumers on online shopping platforms are expressing positive or negative emotion about particular products. Automatically analyzing the sentiment of products can help shopping platforms identify new products with positive reviews that could be highlighted to increase sales, or allow manufacturers to quickly sift through a large number of product reviews to identify potential areas of improvement *the utility of working on the task.*

Currently, there is *gap in knowledge*, for example, lack of sentiment analysis for a particular subdomain (e.g. books about turtles). In this work, we collect a dataset of 200 reviews of turtle books and provide human labels of their sentiment. We then train a XYZ model using this dataset, showing that we can reach an overall performance of XX% for this important task *(how you address the need/the contribution)*. See Figure 1 for an example.

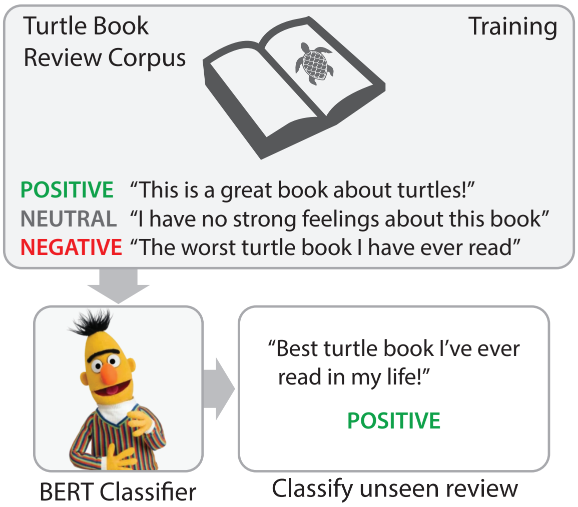

Figure 1: An example of the sentiment analysis task. A corpus of turtle book reviews serves as training data to fine-tune a BERT classification model, trained to classify reviews as *positive, neutral,* or *negative*. The classifier is then evaluated on unseen reviews.

The Introduction is typically 1-1.5 pages (i.e. may spill onto the first column of the second page). The total length of your paper should follow the guideleins for a short ACL paper – 4 pages, plus additional pages for references. 539 students may use 4 or 5 pages.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2 Related Work

In this section, provide a brief review of 3 papers that are related to your task. Normally, this is where you review the state-of-the-art in this subfield or task, and briefly contrast it with your approach. Here, you just need to identify 3 paper that are in the same area, and describe how they frame the task, what methods they use, and their main results. Citations should be entered into the custom.bib file, and cited using the appropriate latex citation commands, e.g. `\cite`.

Aho et al. (1972) provided the first detailed study of sentiment analysis on turtle book reviews, where they use model XYZ on an in-house dataset of 50 turtle book reviews. Compared to Aho et al. (1972), this work includes larger training evaluation, and uses fancy model 2.0 that improves ABC.

Several groups have been interested in sentiment analysis on non-turtle books (e.g. Chandra et al., 1981; Gusfield, 1997). For example, Ando et al. (2005) explore the related task of sentiment analysis on near-domain books such as lizard books, and pamphlets describing other reptiles. etc.

The related work should be approximately half a page.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3 Approach

How do you frame the problem? Classification, ranking, sequence2sequence, etc? Do you need a figure to describe the actual task? What are the important things to know about your task? e.g. if it's classification, how many classes?

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3.1 Data collection

Did you collect and/or label your own data for this task? Describe the collection and labeling procedure here. Provide one or two examples, possibly in a figure.

Did you use someone elses data? Describe it, and provide one or two examples, possibly in a figure.

How is the data divided into train/dev/test sets? (e.g. 50% / 25% / 25%)?

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo.

Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

3.2 Experiments

We evaluate the performance of XX models. Performance is measured using *(How is your task evaluated? Accuracy? F1? Precision@1? etc.).*

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

3.3 Models

If you are doing an experimental paper, you must implement at least two models. One model should be a “baseline” model – a simple model that shows how well a simple method performs on the task. Typically these are n-gram models (e.g. your unigram/bigram model using logistic regression from Assignment 2), ranking using tf.idf vectors, etc.

Model XYZ: Succinctly describe the workings of your baseline model here. What features does it use? What learning framework does it use? etc. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Model	**Precision**	**Recall**	**F1**
Baseline (Unigram+Bigram)	50	40	20
Fancy Model 2.0	60	50	$30^{\dagger}$

Table 1: Task performance across the models under investigation. † signifies that performance is significantly different from baseline performance ($p < 0.05$) using a non-parametric bootstrap resampling test.

Model ABC: Succinctly describe the workings of your other fancy model here. For example, we make use of a BERT-based classifier (Devlin et al., 2018) fine tuned to perform the sentiment analysis task. For computational tractibility we make use of the TinyBERT (Jiao et al., 2020) classifier, which reduces BERT from a 110M parameter model to a XXM parameter model with reduced computational requirements, while retaining much of the performance benefits of the original model. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

3.4 Hyperparameter Tuning

You should report performance on both the **development** and **test** sets in your tables/figures, and describe any hyperparameters that were tuned on the development set.

3.5 Results

The experiments are described in Table 1. Overall the results show...

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3.6 Error Analysis

Students enrolled in 539 must complete an error analysis (and extend their papers into an additional page, to 5 pages). Analyze 50-100 randomly picked errorful predictions that the model makes (on the development set), and distill them into a number of specific categories. Those are shown in Table 2 and described below.

Complex Examples (50%): Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Low frequency words (10%): Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Sarcasm (5%): Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Prop.	**Error Class**
50%	Complex examples with ambiguous words
10%	Unseen data has low-frequency words
5%	Examples contain sarcasm
...	

Table 2: Common error classes and proportions of errors for 100 randomly selected errors on the development set.

4 Conclusion

A short summary paragraph (generally up to one third of one column) summarizing the results. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5 Project Site

The project is available at `http://www.github.com/....`

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

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*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

A Example Appendix

This is an appendix.