NLP Paper title pending

Yasmine Benmessaoud, yabe@itu.dk
Cæcilie Abildgaard Jeppesen, cjep@itu.dk
Luke O'Neill, luon@itu.dk

Abstract

Abstract pending...

1 Introduction

Sentiment analysis of text using machine learning is used as a proxy for human emotional assessment. (Kenyon-Dean et al., 2018)
We are in an increasingly digital age, with an ever increasing amount of online communication. Providing some robust method for analysing the sentiment of text could provide solutions to the challenges faced in online spaces such as bullying (Tika Dewi Amelia and Rania Balqis, 2023) as well as some more business focused tasks.

Historically, when testing a model's success, some straightforward analyses are performed, commonly an accuracy measure. In categorical sentiment analyses, this is occasionally taken a step further, with an F-test to compare category accuracy variance. Simple methods such as these can lead to overestimation or misunderstanding of a model's performance.(Ribeiro et al., 2020)

CheckList is proposed as a method to analyse the nuance of how a model is succeeding or failing, allowing for an analysis that better informs the directions for improvement.
Part of the focus of CheckList is on Named Entity Recognition (NER)

Generally humans would have no preconceived sentiment attached to an unknown Named Entity (NE). When reading text, the linguistic indicators of sentiment would more likely come from adverbs, adjectives and participles. With this understanding, CheckList proposes Invariance (INV) as one of the tests to assess the performance of an NLP model. If a NE is replaced in text with another in such a way that has no syntactic change, does the predicted sentiment of the text change? Examples of how this type of testing works is shown in figure 1.

Test case	Expected	Predicted	Pass?
Testing **NER** with ***INV*** — Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	x
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	x

Figure 1: INV test examples from the CheckList Evaluation paper (Ribeiro et al., 2020)

Conversely, the test of Directional Expectation (DIR), focuses on the ability of adverbs, adjectives and participles to affect a models certainty of sentiment prediction.(Ribeiro et al., 2020)
In short, these two tests, INV and DIR, will look at how a sentiment analysis model responds to changes in NEs, and conversely to changes in language with senitment indication whilst NEs remains unchanged. In this way we can provide detailed analysis of how NEs affect the outcome of a sentiment analysis model.

Furthermore, we aim to study the ability of a sentiment model to correctly predict sentiment from a domain of which it wasn't trained. This was in the hopes that we could simulate the real world challenge of a model being used on new, unseen data. In this way we could counteract the bias created by using only a held-out dataset, but also assess the level of bias contained in our model by comparing in and out of domain tests. Our in-domain is movie reviews from the IMDB dataset(Maas et al., 2011) whilst out-of-domain is a Yelp reviews dataset (Anonymous, 2023)

2 Experimental setup

2.1 Data setup

We use the IMDb movie reviews dataset, 50,000 written reviews of movies and TV shows, categorised as positive or negative sentiment. The two sentiment categories were initially evenly balanced, with 25,000 reviews each.

Generally both categories have similar distributions of word counts as seen in figure 2

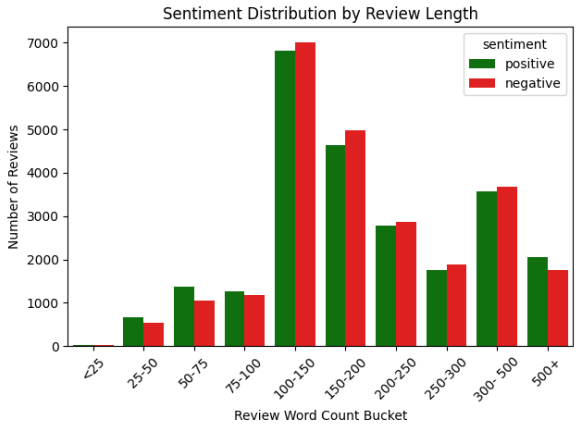

Figure 2: Review length distributions by sentiment category

Data cleaning was mostly motivated by looking at word frequency and searching for anomalous characters. The most common issue being that line breaks had been retained as strings, and so needed removal. Furthermore, we performed reg-ex based identification of punctuation. html tags, urls, and common emojis and then removed these items. (Al Sharou et al., 2021) Our code for this sub-task was inspired by a kaggle project on the same dataset (Abdullah). The last cleaning method we performed was identification and removal of duplicates using the drop_duplicates dataframe method.

Our final cleaned and processed dataset consists of 24,884 positive reviews and 24,697 negative ones.

Additional figures from the EDA process are available in the Appendix.

2.2 Named Entity Recognition Model Setup

2.3 Creating Test Data

In order to perform the CheckList evaluation tests on our in-domain data we chose to hold-out 10% of the reviews. This was in an attempt to utilise the majority of data fro training the sentimental model.

We then performed Invariance perturbation on [half?] of these reviews and Directional Expectation on [the other half?]

2.3.1 Invariance

Having performed NER on our review data, we systematically isolated and replaced the named entities with ones from two json libraries of tagged Named Entities [Cite the libraries after making sure you know which ones they are, and what they're used for]
. This was done within categories, meaning that a PER tagged named entity would be replace with a PER named entity, and likewise for ORG and LOC entities. Entities created from more than one token were also ensured to be replaced with randomly generated entities of equivalent length so that potential random differences according to review length would not be introduced.

2.3.2 Directional Expectation

In order to perform directional expectation, we created [X number] of each positive and negative sentiment short sentences without named entities. According to the ground truth sentiment of a label, [or should this be what the model predicts ??? double check with Checklist] we appended an equivalent sentiment from our list of short sentences to the review.

Then utilising the original unperturbed test sentence and it's new pertubation we assessed the difference in confidence of the model of it's sentimental categorisation. The intention being that, because our added short sentences would contain mostly adverbs, adjectives and participles, and explicityl no named entities, we could accuratley assess whether our model was responding to the types of words that humans also supposedly use to detect sentiment.

3 Sentimental Analysis Model

3.1 Model training

We utilised 90% of our IMDB review data to train our chosen sentimental model, this was in an attempt to ensure the most accuracte model [maybe change if we change train %]

3.2 How it works (given our test data)

4 Results

4.1 Traditional Metrics

According to a traditional, in-domain evaluation metric of accuracy, the model was correct X% of the time.

A break down of the results by sentiment category can be seen in table 1.

Data	Data Split	Accuracy
IMDB test	Total	??%
	positive	??%
	negative	??%
Yelp data	Total	??%
	positive	??%
	negative	??%

Table 1: A table of sentiment accuracy by category.

4.2 CheckList Evaluation tests

4.2.1 Invariance

4.2.2 Directional Expectation

5 Analysis

6 Discussion

7 Limitations

8 Conclusion

References

Sheikh Muhammad Abdullah. Text preprocessing | nlp | steps to process text.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Anonymous. 2023. Opedabsa: A dataset for open domain aspect-based sentiment analysis from public reviews. In *ACL Rolling Review - September 2022: A new initiative of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Lolitha Tika Dewi Amelia and Nadira Rania Balqis. 2023. Changes in communication patterns in the digital age. *ARRUS Journal of Social Sciences and Humanities*, 3(4):544–556.

Group Contributions:

Appendix A. Chart showing sentiment distribution

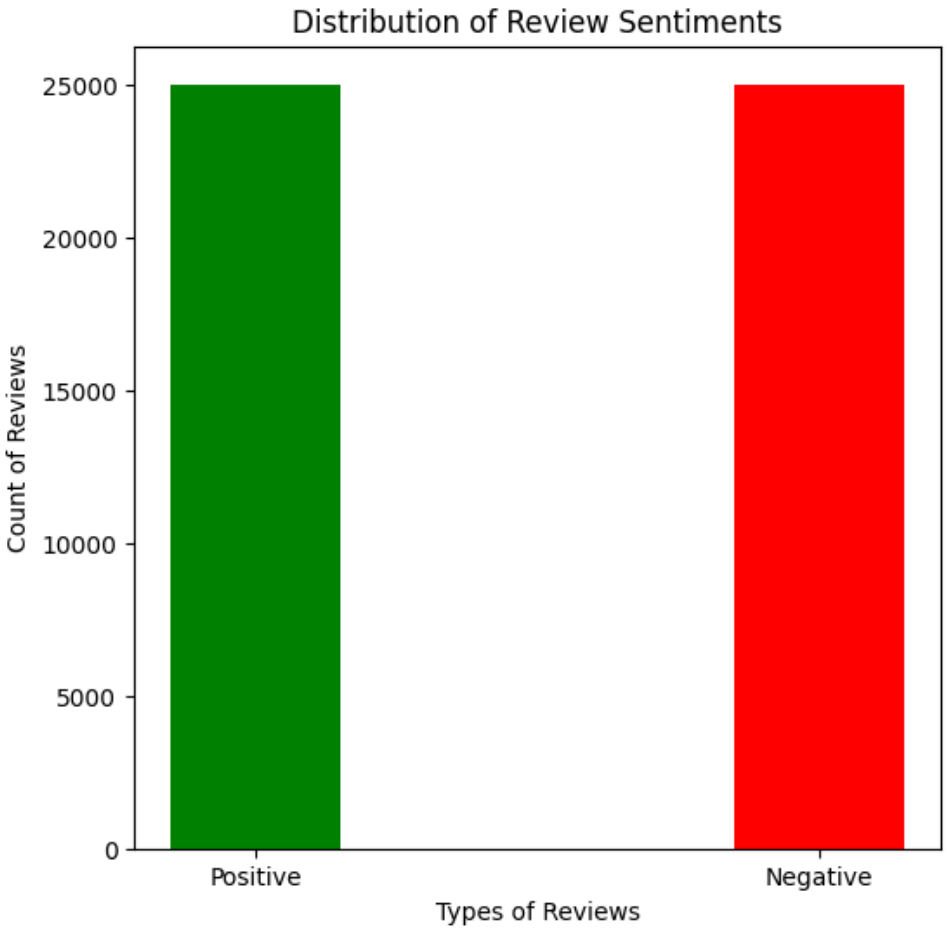

Figure 3: Distribution of Review Lengths

Appendix B. Two charts showing word count distibution by sentiment

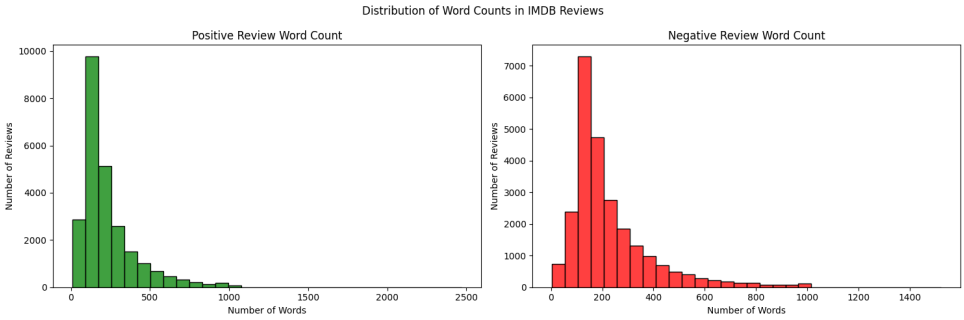

Figure 4: Sentiment Label Distribution

Appendix C. Overlayed word count distributions by sentiment, with bin sizes relative to the review of largest word count

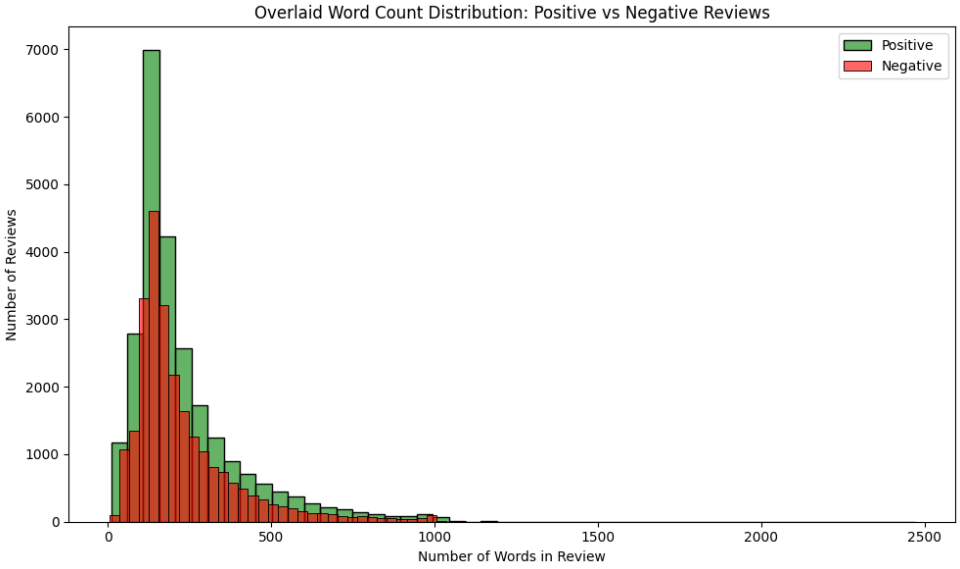

Figure 5: Word Cloud for Positive Reviews

Appendix D. Bar chart of word counts

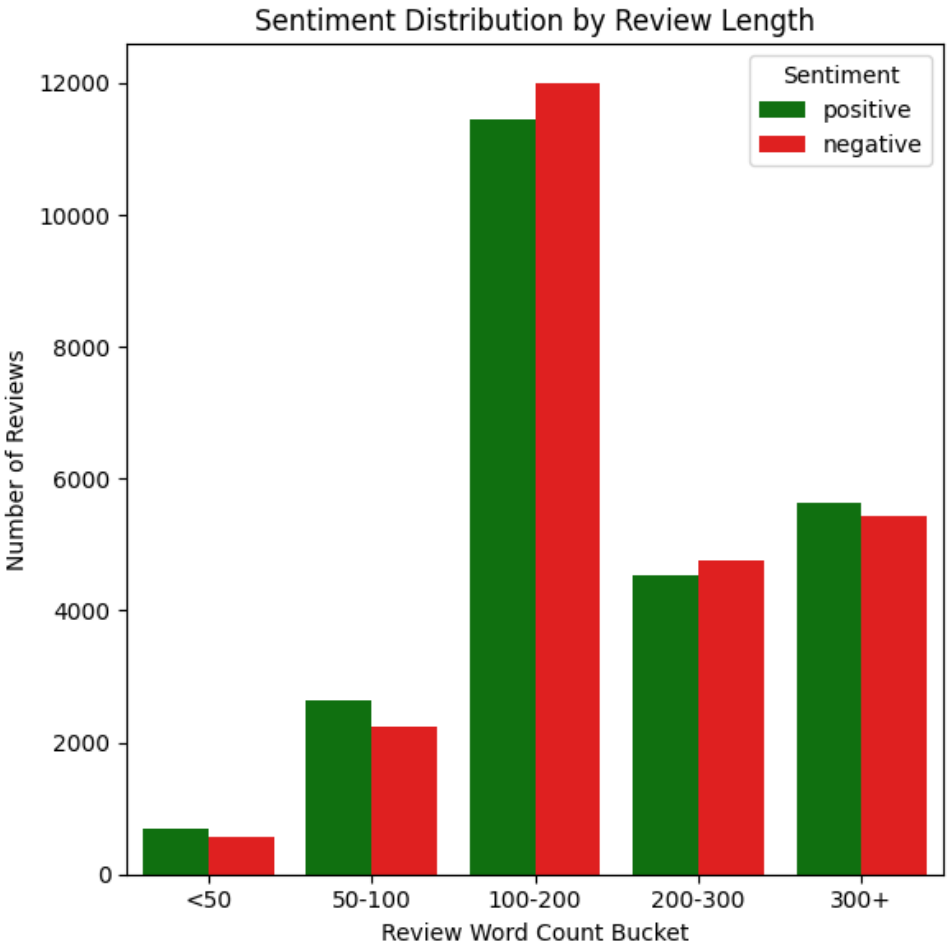

Figure 6: Word Cloud for Negative Reviews

Appendix E. Frequency chart of words by sentiment prior to data cleaning

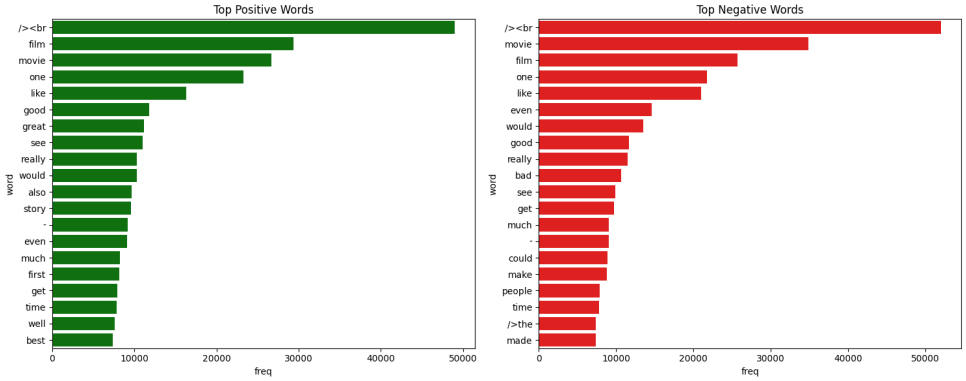

Figure 7: Stopword Frequency Distribution