# Introduction

1. **Introduction to the Domain-Specific Area (max. 500 words)**

* Start with an introduction to the “domain specific area” 🡪 what is the domain-specific area here?
* Identify the **problem or challenge** (an area where text classification methods can **contribute**)
* Include **relevant literature** to show that this area has some **significance**. **How can text classification contribute to this domain?**

**Identify the Domain or Challenge:**

* The domain is *sentiment analysis of poetry* and more generally, the application of computational methods to literary studies.
* Sentiment polarity detection, multi-class classification.
* A challenge because by nature poetic text is full of figurative expressions and indirect descriptions of emotion compared to normal restaurant or shop reviews.

**Associated Dataset:**

* <https://huggingface.co/datasets/google-research-datasets/poem_sentiment>
* “Poem Sentiment is a sentiment dataset of poem verses from Project Gutenberg. This dataset can be used for tasks such as sentiment classification or style transfer for poems.”
* The dataset is already split into training, validation and test sets. Keeping the same split allows for **reproducibility** and **comparability**. It allows others to validate the findings and allows for comparison with benchmarks and SOTA scores. You can use *other splits as well* as long as you justify why, for example, **more robustness** by using **cross-validation** techniques and model’s performance on different **subsets of the data**. It is important though to also test your models on the **same split** to allow for consistency with previous research. Clearly document why you want new splits and the techniques used (cross-validation, random split, stratification). Only change the split if it seems like a reason emerges to do so.
* Language ‘en’
* Each example has an ID/index, the verse\_text, and the sentiment label.
* The original dataset used -1 for negative, 0 for neutral and 1 for positive.
* The sentiment labels here are
* 0 for negative,
* 1 for positive
* 2 for neutral
* 3 mixed (negative AND positive) 🡪 ambiguous!
* **Licensing (Creative Commons 4.0):** <https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/cc-by-4.0.md> 🡪 *permits almost any use subject to providing credit and license notice*.
  + Create a text file (typically named LICENSE or LICENSE.txt) in the root of your source code and copy the text of the license into the file.
  + It is also acceptable to solely supply a link to a copy of the license, usually to the <a href='https://creativecommons.org/licenses/by/4.0/'> canonical URL for the license </a>.
  + This work is licensed under a Creative Commons Attribution 4.0 International License
  + **CC BY 4.0 LICENSE TERMS (**[**https://www.ebi.ac.uk/ols4/ontologies/mcro/classes/http%253A%252F%252Fwww.ebi.ac.uk%252Fswo%252Flicense%252FSWO\_1000065**](https://www.ebi.ac.uk/ols4/ontologies/mcro/classes/http%253A%252F%252Fwww.ebi.ac.uk%252Fswo%252Flicense%252FSWO_1000065)**) 🡺** A non-copyleft license that is good for art + entertainment works + **educational** works. People are free to share, copy and redistribute the material in *any medium or format*, and adapt and transform the material even for commercial purposes. The licensor cannot revoke the freedoms as long as you follow the license terms. The terms are **attribution**, you must **give appropriate credit** and **provide a link to the license** **HERE:** <https://creativecommons.org/licenses/by/4.0/>
* **Attribution:**

**Title of Dataset:** Poem Sentiment Dataset

**Brief Description:** A sentiment dataset of poem verses from Project Gutenberg with annotations tagged as *expert-annotated* on HuggingFace.

**Source:** [HuggingFace Google Research Datasets: *Poem Sentiment Dataset*](https://huggingface.co/datasets/google-research-datasets/poem_sentiment)

**Link on GitHub:** [GitHub poem-sentiment dataset in .tsv files by authors: *duthus* and *socialmkhan*](https://github.com/google-research-datasets/poem-sentiment)

**Authors:** [Google Research Datasets](https://huggingface.co/google-research-datasets) by [Albert Villanova](https://huggingface.co/albertvillanova)

**Contributor (added dataset to HuggingFace):** [@patil-suraj](https://github.com/patil-suraj)

**License:** [**CC BY 4.0**](https://creativecommons.org/licenses/by/4.0/))

* You must also indicate clearly at the end if changes were made in any reasonable manner, not in any way endorses you or your use.
* Citation information:

@misc{sheng2020investigating,

title={Investigating Societal Biases in a Poetry Composition System},

author={Emily Sheng and David Uthus},

year={2020},

eprint={2011.02686},

archivePrefix={arXiv},

primaryClass={cs.CL} }

**“Additionally, your work must be situated within existing literature, with proper citations and references to high-quality sources.”:**

**To get 4/5 marks:** Domain-specific area clearly stated, informative description, fully referenced work.

* **Performance of advanced models on this dataset:**
  + [distilbert-base-uncased-finetuned-rating-poem](https://huggingface.co/VuaCoBac/distilbert-base-uncased-finetuned-rating-poem)
  + Apache 2.0 license 🡪 can use and modify, but must post link to license, attribution and indicate the changes made.
  + A fine-tuned version of distilbert-base-uncased (**look this up!)**
  + Results on evaluation set: **accuracy** of **0.8762** and **F1** of **0.8765** 🡺 difficult to beat! Learning rate was 2e-05, batch size 8, seed 42, Adam optimizer (look at link for more details), linear lr scheduler, trained for 5 epochs. Also look at link for versions of Python packages (e.g. transformers, PyTorch) used.
  + Weird thing about this model (on the site) 🡪 if you type in *I really love him*, it comes out with 0.999 neutral.
  + **Another model:** 
    - [Bert\_uncased\_fine\_tuned\_Reward\_Model](https://huggingface.co/AliChazz/Bert_uncased_fine_tuned_Reward_Model)
    - **Accuracy** on the poem dataset evaluation set is **0.875** (similar to the other one)
    - **Trained for 20 epochs this time** 🡪 did not achieve a better result…
    - Look at link for list of hyperparameters.