# **Notes from Research Articles for Introduction to the Domain**

**Domain Definition**

* Sentiment analysis of poetic texts (text classification) 🡪 a step towards computational emotion analysis in literature/understanding the emotionality and sentiment polarity of poetic texts (<https://zfdg.de/2019_008>)
* An important field in digital humanities, which is “still being formed as a field” (ref: <https://zfdg.de/2019_008>)

**Why this is significant/important**

**Contribution of my paper**

**Article 1: *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies* (Evgeny Kim and Roman Klinger), 2021, DOI:** [**10.17175/2019\_008\_v2**](http://dx.doi.org/10.17175/2019_008_v2) **Link:** [**https://zfdg.de/2019\_008**](https://zfdg.de/2019_008) **from “Zeitschrift fuer digitale Geisteswissenschaften”/”Journal of Digital Humanities”**

* Sentiment analysis as part of **computational literary studies**.
* Following Liu: “we define sentiment as a *positive* or *negative* feeling underlying the opinion”. Another interpretation of *sentiment analysis*: “a broader description of a research field, which considers affective computing applied to textual analysis”, including the distinction between subjective and objective statements.
* Defining the concept of “emotion” is a very difficult task – involves a set of expressive/behavioural/physiological/phenomenological features. “***In this view, an emotion can be defined as »an integrated feeling state involving physiological changes, motor-preparedness, cognitions about action, and inner experiences that emerges from an appraisal of the self or situation«”***
* Sentiment and emotions can be analyzed computationally. This work **does not focus on binary polarities**, but on emotion recognition.
* **Computational Literary Studies:** Da defines comp literary studies as ***“the statistical representation of patterns discovered in text mining fitted to currently existing knowledge about literature, literary history, and textual production*.” Refers to the *“practice of running a textual analysis on a computer to yield quantitative results”*.**
* **Approaches to emotion/sentiment analysis in literature:**
  + Lexicon-based methods 🡪 in digital humanities, where the object of research/corpus is of higher importance, annotators need to be domain experts for a particular object of research, comes at the cost of accuracy.
  + Feature-based ML/traditional stats techniques such as SVMs using word frequencies, n-grams (sequences of adjacent words), or other **linguistic features** to predict labels (emotions), or random forest.
  + Representation-learning/DL-based: RNNs (to recognize patterns in sequences one element at a time, with hidden states summarizing the sequence up until that point), LSTM (captures long-term dependencies), transformer models (SOTA) such as BERT, these DL techniques capture complex contextual relationships and dependencies.
* **Why is sentiment analysis of poetic/literary texts important?**
  + Emotions influence the experience of our daily life.
  + A growing body of research pinpoints the importance of emotions for **literary comprehension**, people make deliberate choices with regard to their emotional states when seeking books/movies (applications: recommendation engines, product advertising)
  + The “digital revolution”/digital humanities due to new methods of quantitative research, not all methods yet exploited 🡪 there is so much literary/text data and these methods are relatively new. Humanities scholars make use of **new tools of text analytics and data-driven approaches to theory formulation (or corroboration).**
  + **1982:** the first work on computer-assisted emotion modelling in literature (Anderso n and McMaster) 🡪 “the emotional tone of a collection of texts can be responsible for the reader’s interest”, and the results of the study suggest that a large-scale analysis of the emotional tone of a “collection of texts” is possible with computer programs
  + **Main implications of Anderson’s and McMasters’ findings:** identifying the emotional tone of text passages can model the **affective patterns** of a text/collection of texts, which can in turn be used **to challenge/test existing literary theories**. Secondly, their work demonstrates that the **stylistic properties of texts can be defined on the basis of their emotional interest**, and **not only** on their linguistic characteristics. A big area now in the digital humanities is sentiment analysis in literature.
  + **In recent years the “affective turn”** **has been taken by many disciplines.**
  + **A key challenge:** how emotions are communicated clearly point to challenges for computational methods 🡪 implicit descriptions, “world knowledge” and “inference steps” grounded in the interplay between the text and readers’ experiences 🡺 not tackled with computational methods yet! E.g. Kafka = non-verbal cues described as indicators of emotional states, very implicit/depending on **world knowledge**.
  + **Another challenge:** usually short units of text are used, lack of context for character development (novels), prediction depends on the **local** description (**ADDRESS THIS SHORTCOMING IN THE EVALUATION!)** a clear disadvantage of current computational linguistics.
  + **Linguistic features important for emotion detection:** Kuivalainen on Katherine Mansfield’s prose 🡪 use of adjectives, adverbs, **deictic markers** (to pinpoint the source of emotions), orthography, shifts between psycho-narrative and free indirect discourse, internal discourse is marked by dashes, exclamation marks (do not remove punctuation!), intensifiers, repetition= trigger an emotional climax. Emotion-bearing words are only sometimes used – a **specific challenge for classifying poetic text** – due to many writers striving to express emotions by way of figures of speech, lots of ambiguity, showing rather than telling, “catachresis” = using a word in an unconventional manner to stretch its original meaning/use it figuratively, creating a new, unusual expression = very important for poets, but very difficult for machines! E.g. “he turned to the bottle” (i.e. alcohol). Unconventional imagery and language.
  + **KEY CHALLENGES:** poetry is laden with *indirect* representation of emotion. Conrad landscape example. Indirect portrayal, characters’ nonverbal behaviour. Problem of text fragments being too short/lacking context.
  + “It seems that one should search for a balance between low-level linguistic feature analysis of emotional language and a rigorous high-level hermeneutic inquiry dissecting the form of the novel and its under-covered philosophical layers”
  + Classing books or poems on whether there is a happy ending 🡪 important for recommendation systems!
  + Barros et al: classifying 185 Francisco de Quevedo’s poems into emotional categories by constructing a **lexicon of emotion words** and translating them into Spanish, a task-specific lexicon, to which each poem is then compared **based on normalized term counts** 🡪 decision trees then used as well as **resampling the collection**, success was achieved.
  + Ethan Reed: dictionary-based black-box sentiment analysis outputting the polarity of 20th century American poetry (60s and 70s *Black Arts Movement*), how feelings of injustice are coded in terms of race and gender, what sentiment analysis can show us about the relations between affect and gender in poetry. **Reed notes that the surface value of a word can sometimes contradict the more nuanced meaning it has in a social and political context.**
  + Yu: ranking emotionality of sentimentalist early American novels = noteworthy due to the larger **unit of analysis**. Ranking texts as *highly emotional* or the opposite using SVMs and Naïve Bayes. **The results of the evaluation suggest that arbitrary feature reduction steps such as stemming and stopword removal should be taken very carefully, as they may affect the prediction.**
  + Volkova: problems with annotation, whether a word is neg/pos can rely on social context etc. Problems of granularity, texts being too short.
  + Sentiment **polarity** is what I’m doing here.
  + Ashok et al: sentiment polarity to predict the success of a book (these approaches are often criticized), found that unsuccessful stories contained **more discriminative words that had a negative connotation**.
  + Zehe et al and happy ending classification: ***“A novel is considered to have a happy ending if the situation of the main characters in the novel improves towards the end or is constantly favorable. The novels were manually annotated with this information by domain experts. For feature extraction, the authors first split each novel into n segments of the same length. They then calculate sentiment values for each of the segments based on a normalized word frequency with a German version of the NRC Word-Emotion Association Lexicon.***[***[60]***](https://zfdg.de/2019_008#fn60)***An automatic sentiment classification with support vector machines achieves reasonable and encouraging results.”***
  + Sentiment/emotion features can be used as features for **higher-level classification**, such as genre classification 🡪 different literary genres = different emotions, but it is **not always that straightforward and reliable**.
  + Reagan et al: splitting novels into equal-length sections and computing happiness scores for each to analyze narrative patterns, and find which patterns are more popular with readers based on download counts. Useful to construct higher level vectors based on emotions for genre classification.
  + **Challenge:** negation!
  + **Sentiment analysis can be used for analyzing political preferences of the electorate or for mining opinions about different products or topics.**
  + **Similarly, several digital humanities studies incorporate sentiment analysis methods in a task of mining sentiments and emotions of people who lived in the past. The goal of these studies is not only to recognize sentiments, but also to understand how they were formed.**

**Super Interesting: Topography of Emotions**

* + Heuser et al.[[71]](https://zfdg.de/2019_008#fn71) start with a premise that emotions occur at a specific moment in time and space, thus making it possible to link emotions to specific geographical locations. Consequently, having such information at hand, one can understand which emotions are hidden behind certain landmarks. As a proof-of-concept, Heuser et al. build an [interactive map of emotions](https://www.historypin.org/en/victorian-london/) in Victorian London[[72]](https://zfdg.de/2019_008#fn72) where each location is tagged with emotion labels. The underlying corpus for their analysis consists of English books from the eighteenth and nineteenth century, from which they extract frequently mentioned geographical locations of London. The presegmented data is then given to annotators who are asked to define whether each of the passages expressed happiness or fear, or neutrality. The same data is further analyzed with a dictionary-based sentiment classifier.
  + [[50](https://zfdg.de/2019_008#pid50)]Some striking observations are made with regard to the data analysis. First, there is a clear discrepancy between fiction and reality – while toponyms from the West End with Westminster and the City are over-represented in the books, the same does not hold true for the East End with Tower Hamlets, Southwark, and Hackney. Hence, there is less information about emotions pertaining to these particular London locations. Another striking detail is that the resulting map is dominated by the neutral emotion. Heuser et al. argue that this has nothing to do with the absence of emotions but rather stems from the fact that emotions tend to be silenced in public domain, which influenced the annotators decision.
  + [[51](https://zfdg.de/2019_008#pid51)]The space and time context are also used by Bruggman and Fabrikant[[73]](https://zfdg.de/2019_008#fn73) who model sentiments of Swiss historians towards places in Switzerland in different historical periods. As the authors note, it is unlikely that a historian will directly express attitudes towards certain toponyms, but it is very likely that words they use to describe those can bear some negative connotation (e.g. cholera, death). Correspondingly, such places should be identified as bearing negative sentiment by a sentiment analysis tool. Additionally, they study the changes of sentiment towards a particular place over time. Using the General Inquirer (GI) lexicon[[74]](https://zfdg.de/2019_008#fn74) to identify positive and negative terms in the document, they assign sentiment scores and conclude that the results of their analysis look promising, especially regarding negatively scored articles.
  + **Other papers in this category link sentiment and emotion to certain groups, rather than geographical locations. The goal of these studies is to understand how sentiment within and towards these groups was formed.** E.g. tracking an author’s reputation.
  + **Another challenge (for historical texts):** The overall conclusion of their work is that the assignment of a polarity in the historical domain is a challenging task largely due to lack of agreement on polarity of historical sources between human annotators. Different affective lexicons, really need expert knowledge.
  + **Irony and deceit play a crucial role in many literary works.**
  + **Morin and Acerbi remind us that the Romantic period was dominated by emotionality in writing, which could be the effect of a group of writers who wrote above the mean. If one assumes that each new writer tends to copy the emotional style of their predecessors, then writers at one point of time are disproportionally influenced by this group of above-the-mean writers. However, this trend does not last forever and, sooner or later, the trend reverts to the mean, as each writer reverts to a normal level of emotionality.**
  + “We have shown throughout this survey that there is a growing interest in sentiment and emotion analysis within computational literary studies as one main field of digital humanities.”
  + “Given the fact that DH have emerged into a thriving science within the past decade, it may safely be said that this direction of research is relatively new. It further constitutes an interesting field that connects literary studies and computational linguistics.”
  + “Moreover, a recent meta-study by Mäntylä et al.[[130]](https://zfdg.de/2019_008#fn130) shows that the number of papers in sentiment analysis is rapidly increasing each year. Indeed, the topic has not yet outrun itself and we should not expect to see it vanishing within the next decade or two. In addition, there are still many open challenges.”

**Important Point for Why to Use Traditional Statistical Methods instead of DL:**

* “*Transparency of the computational method is not a bonus; it is a crucial property.* In digital humanities, research is often exploratory. The application of an existing method on a corpus can lead to new findings, but it is common that an interactive application of a method to explore a phenomenon is even more promising. Such interactive application requires full control by the user in real time – and that is something that pretrained deep neural methods cannot (yet) provide. However, emotion lexicons that point to particular aspects in the text in a transparent manner do, despite of their disadvantages.”
* In contrast to most emotion analysis work in other domains (like social media or news), the unit of analysis should be larger. It is not sufficient to only analyze sentences in isolation (or even just words). Instead, the overall development of characters, the story line as a whole need to be considered. This is a research direction that hardly received any attention yet; presumably because of technical challenges, but likely also due to the lack of annotated corpora that would be required to contain annotations on different levels. Further, these annotations need particular expertise from the annotators. It is not feasible to show an entire book to workers on a crowdsourcing platform to receive annotations on fine-grained levels (for characters and their developments). Therefore, for domains of interest, we point out that the development of corpora in computational literary studies are expected to be more expensive and will take longer than in other fields in which emotion analysis is applied.
* While on Twitter analysis, we typically care about the emotion that the author of a message felt while writing it, we typically do not care about the emotion of the author of a novel, while writing it.[[132]](https://zfdg.de/2019_008#fn132) Instead, we are faced with the more challenging task to attribute emotions to characters or even infer the emotions that might be developed by readers of a text.
* In summary, we believe that the field of emotion analysis for literary studies has still space for research in multiple directions.

**Article 2: The Computational Case against Computational Literary Studies by Nan Z. Da**

* “Computational literary criticism is prone to fallacious overclaims or misinterpretations of statistical results because it often places itself in a position of making claims based purely on word frequencies without regard to position, syntax, context, and semantics. Word frequencies and the measurement of their differences over time or between works are asked to do an enormous amount of work, standing in for vastly different things.”
* “The Bayes rule is a widely used rule that with every new observation updates the probability distribution; the system is “naïve” because the features are supposed to be independent of one another. You do not tell the algorithm the exact criteria by which to make its classificatory decisions; you tell it what to pay attention to and it learns the decision rule based on some basic features, changing the probability distribution every time a new thing comes in and so getting smarter and better at classifying the next thing.”
* “No one has ever said, though, that consistent word frequency is what distinguishes Shakespeare’s comedies from tragedies, tragedies from histories, and so on—and no one would ever say that because such distinctions cannot be captured with word frequencies.”
* “The hitch of using textual pattern mining for forensic stylometry is that even if you apply pattern recognition techniques that reduce noise and nonlinear interactions between data, the stylistic differences that can be captured for literature tend to be driven by stop words—if, but, and, the, of. 39”
* “Why is that the case? Mark Algee-Hewitt and Piper tell us that “**stop words are usually semantically poor and yet stylistically rich**... . **The best means so far for determining authorship attribution and classifying texts as categorically different**.”40
* In reality, stylistic differences boiling down to stop words is not surprising at all.
* To locate a statistical difference of occurrence means having enough things to compare in the first place.
* If the word cake only occurs once in one text and four times in another, there’s no way to really compare them, statistically. **By the numbers, stop words are the words that texts have most in common with one another, which is why their differentiated patterns of use will yield the readiest statistical differences and why they have to be removed for text mining.”  
  p.624:**
* **Token-Type Ratio and SOC Prediction**:
* Long and So argue that they can predict SOC literature with high accuracy using linguistic traits, including token-type ratio (the ratio of total words to unique words).
* They found that SOC passages tend to have a lower token-type ratio, suggesting a higher repetition of words compared to realist literature.
* Impact of Stop Words:
* Stop words are common words that are often removed in text processing tasks because they are frequent and may not carry significant semantic meaning.
* Long and So initially found that when using their own curated stop word list (which included nonstandard stop words), SOC texts tended to have a higher token-type ratio compared to realist texts.
* This led them to conclude that token-type ratio was a strong predictor of SOC.
* Standard Stop Word List vs. Custom Stop Word List:
* The key finding was that when they re-ran their analysis using a standard stop word list (a more conventional set of stop words), the results reversed.
* With the standard stop word list, realist texts showed a higher token-type ratio than SOC texts.
* This reversal occurred because the custom stop word list they initially used was biased towards SOC texts, as SOC stop words tended to be more similar and thus removed less frequently.
* Implications:
* The reliance on a custom stop word list biased their results towards SOC texts.
* Using a standard stop word list revealed that token-type ratio alone may not be as robust a predictor of SOC as initially claimed.
* This highlights the importance of careful preprocessing steps and the potential biases introduced by using nonstandard linguistic features or preprocessing methods.
* In summary, the problem with removing or not removing stop words in the context of predicting SOC lies in the bias introduced by the choice of stop word list. Long and So initially found a strong correlation between token-type ratio (affected by stop word removal) and SOC using their custom list. However, using a standard stop word list revealed that this correlation was due to the custom list's bias, suggesting that token-type ratio alone may not be the definitive indicator of SOC literature.
* **In other sectors and applications, texts with stop words removed can further be categorized—into economic terms, political terms, female consumer, for example. Another level of simple and accurate-enough classification has to occur so that categories can be compared rather than an individual word’s frequencies—this is what allows for the statistical analysis of words. When CLS tries to do this for literature, using various methods to reduce large corpora of words to sensible groupings, it realizes that after the necessary dimensionality reduction is performed—uncommon words taken out, stop words removed, groups of words vectorized to become single points in space—it’s left with only a small portion of what it was originally purporting to study, and these are corralled into groupings so general as to preclude meaningful interpretations.**
* Main point: word counts/frequencies might not be the most important thing when trying to discern differences between documents/texts, not enough context in traditional statistical methods over the whole text.
* “—then whatever the Naïve-Bayes classifier classifies as an English haiku is, by their very definition, an English haiku, as they don’t have a rigorous definition to begin with. B” 🡪 subjectivity of definitions/annotations.
* Main criticisms:
  + ***Da argues that computational approaches often prioritize quantitative metrics over deep literary understanding. Metrics like word frequencies, sentence lengths, or syntactic patterns may provide data, but they risk oversimplifying or distorting the richness and complexity of literary texts.***
  + ***Literary analysis traditionally values interpretative flexibility, where multiple readings and subjective interpretations coexist. Computational methods, on the other hand, tend to favor deterministic outcomes and may overlook the subjective and interpretative aspects of literary criticism.***
  + ***Not enough attention paid to context (usually only bi or at most trigrams are considered, or simple word counts)***.
  + ***Lack of real world, historical and social context is an issue in this field, for understanding themes, motifs, references and literary techniques within their historical context.***

**Book 3: Sentiment Analysis and Opinion Mining by Bing Liu (2012)**

* Sentiment analysis = a large problem space.
* Opinions are key influencers of human behaviours.
* Businesses want to find consumer/public opinions about their products and services.
* “Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies.”
* Explosion of text data, opinions and reviews on the Web, surveys etc.
* “Each site typically contains a huge volume of opinion text that is not always easily deciphered in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are thus needed.”
* Predicting box-office revenues for movies
* Twitter moods to predict the stock market (p. 10)
* Studying opinions on the Web 🡺 predicting political change
* “Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis.”
* “A sentence containing sentiment words may not express any sentiment. This phenomenon happens frequently in several types of sentences. Question (interrogative) sentences and conditional sentences are two important types, e.g., “Can you tell me which Sony camera is good?” and “If I can find a good camera in the shop, I will buy it.” Both these sentences contain the sentiment word “good”, but neither expresses a positive or negative opinion on any specific camera. However, not all conditional sentences or interrogative sentences express no sentiments,” 🡪 importance of punctuation being left in!
* Sarcasm problem
* “Many sentences without sentiment words can also imply opinions. Many of these sentences are actually objective sentences that are used to express some factual information. Again, there are many types of such sentences. Here we just give two examples. The sentence “This washer uses a lot of water” implies a negative sentiment about the washer since it uses a lot of resource (water). The sentence “After sleeping on the mattress for two days, a valley has formed in the middle” expresses a negative opinion about the mattress. This sentence is objective as it states a fact. All these sentences have no sentiment words. These issues all present major challenges. In fact, these are just”
* Negation handling, word sense disambiguation 🡺 common NLP problems.
* “Positive, negative and neutral are called sentiment (or opinion) orientations (or polarities).”
* “Sentiment classification is essentially a text classification problem. Traditional text classification mainly classifies documents of different topics, e.g., politics, sciences, and sports. In such classifications, topicrelated words are the key features. However, in sentiment classification, sentiment or opinion words that indicate positive or negative opinions are more important, e.g., great, excellent, amazing, horrible, bad, worst, etc. Since it is a text classification problem, any existing supervised learning method can be applied, e.g., naïve Bayes classification, and support vector machines (SVM) (Joachims, 1999; Shawe-Taylor and Cristianini, 2000). Pang, Lee and Vaithyanathan (2002) was the first paper to take this approach to classify movie reviews into two classes, positive and negative. It was shown that using unigrams (a bag of words) as features in classification performed quite well with either naïve Bayes or SVM, although the authors also tried a number of other feature options”
* “Like other supervised machine learning applications, **the key for sentiment classification is the engineering of a set of effective features.** Some of the example features are:
  + **Terms and their frequency.** These features are individual words (unigram) and their **n-grams with associated frequency count**s. They are also the most common features used in traditional topic-based text classification. In some cases, word positions may also be considered. **The TF-IDF weighting scheme from information retrieval may be applied too.** As in traditional text classification, these features have been shown highly effective for sentiment classification as well.”
  + **Part of speech.** The part-of-speech (POS) of each word can be important too. Words of different parts of speech (POS) may be treated differently. For example, it was shown that adjectives are important indicators of opinions. Thus, some researchers treated adjectives as special features. However, one can also use all POS tags and their n-grams as features. Note that in this book, we use the standard Penn Treebank POS Tags as shown in Table 3.1 (Santorini, 1990). The Penn Treebank site is at http://www.cis.upenn.edu/ ~treebank/home.html.
  + **Sentiment words and phrases.** Sentiment words are words in a language that are used to express positive or negative sentiments. For example, good, wonderful, and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Most sentiment words are adjectives and adverbs, but nouns (e.g., rubbish, junk, and crap) and verbs (e.g., hate and love) can also be used to express sentiments. Apart from individual words, there are also sentiment phrases and idioms, e.g., cost someone an arm and a leg.
  + **Sentiment shifters.** These are expressions that are used to change the sentiment orientations, e.g., from positive to negative or vice versa. Negation words are the most important class of sentiment shifters. For example, the sentence “I don’t like this camera” is negative. There are also several other types of sentiment shifters. We will discuss them in Section 5.2 too. Such shifters also need to be handled with care because not all occurrences of such words mean sentiment changes. For example, “not” in “not only … but also” does not change sentiment orientation. Syntactic dependency. Words dependency-based features generated from parsing or dependency trees are also tried by researchers.

**Article 4: Investigating Societal Biases in a Poetry Composition System by Emily Sheng and David Uthus (2020)**

* Sentiment analysis can help examine bias (negative sentiment) towards certain geographical locations, demographic groups etc. Important ethical connotations. Here: mitigating bias in poetry composition systems (machine-learning suggested novel verses in the style of classic American poets). “As creative works are often shaped by the lived experiences and timely issues of the creator’s life, a poetry composition system trained on poems from different authors of different eras may reflect a variety of societal biases.” E.g. gender, race bias.
* “Social bias” is defined as unequal social perceptions of different socially-defined groups of people, reinforces representational harms (e.g. stereotypes) and potential allocation harms (e.g. unequal job opportunities).
* The authors examined the language polarity (sentiment) of the suggested verses when different demographic groups are mentioned in the user input. “We use sentiment as a proxy metric for the social perception of demographic groups.”
* A system that **favours different sentiment verse suggestions** for mentions of different demographic groups (e.g. *positive* for demographic A and *negative* for demographic B) could propagate negative associations, amplifying existing demographic inequalities.
* Focusing on negative and positive examples for the user input *The women* and *men*, showed more positive suggested next verses for men and more negative results for the women.
* “Thus, we propose a technique to mitigate biases by making the verses suggested by the poetry system less negative in sentiment”
* “Results show that our method has promising results for both reducing negative verses and keeping the distribution of verse sentiments across groups comparable.”
* “. For the first part of the pipeline, we introduce a poetry sentiment dataset and build a BERT-based (Devlin et al., 2019) sentiment analyzer for poetry”
* The sentiment tools are then used to train a style transfer model to augment data to train the next verse prediction component in the poetry composition system.
* “Specifically, we can influence the model to suggest verses with more positive sentiment while keeping the suggested verse quality comparable. This exploratory study introduces the capabilities of style transfer augmentation to mitigate biases and is an example of how bias mitigation can be applied to creative language tasks and information retrieval components.”
* The tool that was improved “Users can either directly use verse suggestions provided by the system, modify the suggestions, or create their own verses. The suggested verses are generated in the style of various classic American poets (e.g., Walt Whitman, Emily Dickinson).”
* Next verse prediction task/retrieval settings 🡪 little info on bias.
* “we largely treat verse generation as a black box component”
* Used transformer-based models and feed-forward networks, a BERT sentiment classifier.
* “Data augmentation is a technique used to increase the diversity and size of a dataset without actually collecting new data. It involves creating modified versions of existing data points to provide more training examples for a machine learning model. This helps improve the model's generalization capabilities and can be particularly useful when dealing with limited data.”
* “By using style transfer for data augmentation instead of filtering out negative examples, we can circumvent data sparsity issues and promote model robustness.”
* “In the specific context of the provided text, **sentiment style transfer** is mentioned as a data augmentation technique. This involves altering the sentiment of text data (e.g., changing a sentence from negative to positive sentiment) to create new training examples. The goal is to balance the sentiment distribution and mitigate biases that might exist in the training data towards different demographic groups.”
* Style transfer model based on the sentiment classifier.
* **The Dataset (the poem sentiment one I am using):** there was no existing public poetry dataset with sentiment annotations, so two professional annotators labelled randomly picked verses from the Gutenberg Poem Dataset. For each sample, annotators could describe the language as *negative, no impact (neutral), positive, mixed (both neg and pos), does not make sense*. The inter-annotator agreement was 0.53 (Cohen’s kappa) for 1550 annotated samples
* ***“If we remove samples where either annotator chose mixed or does not make sense, the kappa score increases to 0.58. Spearman’s correlation for the samples with labels in the three sentiment categories (negative = -1, no impact = 0, positive = 1) is 0.67. These correlations indicate decently strong inter-annotator agreement. For all annotated samples, we only keep the sample if there is agreement across both annotators and if the label is negative, no impact, or positive.”*** 0.53 Cohen’s kappa indicates **moderate agreement** between annotators. Spearman’s correlation coefficient does not assume the relationship is linear (unlike Pearson’s correlation coefficient), 0.67 = moderately strong correlation. Not perfect though 🡪 some level of subjectivity/difficulty with the annotation task.
* ***“***The decision to only keep samples with agreement across both annotators (and with clear labels) is likely aimed at improving the quality and reliability of the dataset by ensuring that the retained samples are those where the annotators most consistently agreed.” 🡪 model is more robust.
* **THE BENCHMARK/STATE-OF-THE-ART RESULT: *“We fine-tune a pretrained BERT model on the filtered annotated sentiment dataset. We use the uncased version of BERT base with a batch size of 32, learning rate of 1 × 10−5 , maximum sequence length of 128, warmup proportion of 0.1, and train for 5 epochs. The resulting sentiment classifier has a development set accuracy of 85.7% and a test set accuracy of 84.6%.”***
* *“*"Uncased BERT" refers to a version of the BERT (Bidirectional Encoder Representations from Transformers) model where all text is converted to lowercase before being processed by the model. This means that the model does not differentiate between uppercase and lowercase letters. For example, "Cat" and "cat" would be treated as the same token.” 🡪 all lowercase means robust to variations in text casing, case information might not be relevant, simplifies vocabulary, reduces the complexity of training.
* **“**When training a DRG model with this dataset, the model learns to convert negative verses to positive verses and vice versa, though we only use negative to positive conversions for our data augmentation method.”
* **“**The model treats the original pair ("by the path an indian sat", "then i cried and ran away") as a negative example because it contains a negative sentiment in the next verse.”
* **“**The model uses the new pair ("by the path an indian sat", "then i sing that human delight") as a positive example because the sentiment of the next verse has been transferred to positive.”
* **“**Although biases in creative language applications are underexplored, it is important to examine biases in these applications that are primarily intended for social use. Our results indicate that style transfer has potential as an augmentation technique to reduce societal biases.”

**Article 5: A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry by Justine Kao and Dan Jurafsky (2012)**

* “These studies showed that computational methods can reveal interesting statistical properties in poetic language that allow us to better understand and categorize great works of literature (Fabb, 2006).”
* What makes one poem more aesthetically pleasing than another?
* Phonetic features, word choice, balance between order and complexity. Aristotle – a balance between ordinary and unusual words.
* Poetry is “usually intentionally ambiguous and often packs several meanings into a compact passage”, so more difficult.
* **Features of poems:** word frequency (logs), type-token ratio (lexical diversity), less repetition of the same words = more professional poems, is higher type-token ratio correlated with higher aesthetic value and more successful reputation of the poet? Alliteration count, assonance, rhymes.
* **Sentiment and Affect:** do good poets explore the emotional world with more intensity? Used several **sentiment lexicons** for this, e.g. Harvard General Inquirer, 182 word categories including basic sentiment categories. “Another sentiment lexicon is the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). While the General Inquirer was designed for content analysis, LIWC was designed to facilitate the understanding of individuals’ cognitive and emotional states through text analysis. As a result, most of the categories in LIWC involve mental activity, with over 4, 500 words related to affective, social, and cognitive processes. Six categories from the Harvard General Inquirer and two categories from LIWC were selected because they are most suitable for our purpose of analyzing elements of poetic craft.”
* **“Show don’t tell”** 🡪 another difficult thing with categorizing poetic emotion! Effective imagery allows readers to bring in their own associations to understand and truly experience a new emotion, and skilled poets and writers are able to pick out specific sensory details that evoke deeper abstractions and generalizations. Described as a “sacred rule”.
* “Another reason why imagery is an essential element of poetic craft is that it allows writers to avoid falling into cliche, which is the bane of the creative writer’s existence.”
* Used logistic regression to categorize poems into poems by a famous vs an amateur poet. Main predictors: type-token ratio, rhyme and alliteration stuff, **fewer negative emotional words**.
* “The fact that professional poets are significantly less likely to use explicitly negative emotion words than amateur poets, but not significantly less likely to use negatively connotative words, suggests that professional poets may evoke more negative sentiment through connotation rather than explicit descriptions.”