The Stylometry of Maoism: Quantifying the Language of Mao Zedong

Maciej Kurzynski
Lingnan University, Hong Kong
maciej.kurzynski@ln.edu.hk

Abstract

Recent advances in computational stylometry have enabled scholars to detect authorial signals with a high degree of precision, but the focus on accuracy comes at the expense of explainability: powerful black-box models are often of little use to traditional humanistic disciplines. With this in mind, we have conducted stylometric experiments on Maospeak, a language style shaped by the writings and speeches of Mao Zedong. We measure per-token perplexity across different GPT models, compute Kullback–Leibler divergences between local and global vocabulary distributions, and train a TF-IDF classifier to examine how the modern Chinese language has been transformed to convey the tenets of Maoist doctrine. We offer a computational interpretation of ideology as reduction in perplexity and increase in systematicity of language use.

1 Introduction

Stylometry, the quantitative analysis of literary style, has been extensively used to study various authors' writing styles, leveraging linguistic features such as word frequencies, sentence length, and syntactic patterns (Stamatatos, 2009; Neal et al., 2017). Historically, stylometry dates back to analyses of narrative style in Shakespeare (Burrows, 1987) and attempts to identify the authors of the disputed Federalist Papers (Mosteller and Wallace, 1963; Tweedie et al., 1996). The advent of computational methodologies has significantly enhanced the scope and depth of stylometric analyses, allowing for the examination of larger textual corpora and the incorporation of multifaceted linguistic features, such as lexical richness, syntactic complexity, and semantic coherence (Seroussi et al., 2014; Sari et al., 2018). Computational stylometry has evolved to include a range of techniques, from Principal Component Analysis (PCA) to various machine learning algorithms and large language models (LLMs), enhancing the reliability of stylistic differentiation (Ruder et al., 2016; Ou et al., 2023).

However, the focus on precision metrics in authorship attribution comes at the expense of interpretability and applicability in humanistic research and teaching. The fact that a particular author uses definite articles or certain grammatical particles more often than others might shed light on their subconscious stylistic preferences, and it might even be decisive in distinguishing their stylistic fingerprints, but similar analyses can hardly explain why a given language has been particularly successful at conveying political messages and furthering domestic mobilization, as was the case of Maoism in post-1949 China.

2 Quantifying Maospeak

Maospeak, Mao-style prose, or *maowenti* (毛文体, also called *maoyu* 毛语), is a set of stylistic features influenced by the writings and speeches of Mao Zedong (1893-1976), the leader of the Chinese communist revolution (Li, 1998). Maospeak has had a transformative impact on the way people in China express themselves, and it continues to affect the everyday language in the PRC even today. Consider the following examples:

- 共产党内部也有斗争。不斗争就不能进步, 不和平。八亿人口, 不斗行吗?

  There is also [class] struggle within the Communist Party. Without struggle, there can be no progress, no peace. With a population of 800 million, how can we not struggle?

- 在我们的面前有两类社会矛盾，这就是敌我之间的矛盾和人民内部的矛盾。这是性质完全不同的两类矛盾。

  There are two types of social contradictions in front of us: the contradictions between ourselves and our enemies and the contradictions within the people. These are two types of contradictions with completely different natures.

Such repetitive, redundant, and depersonalized sentences are a staple of the *Little Red Book*, a compilation of Mao's quotations and a condensed example of the Maoist prose. Despite its thematic coherence, however, Maospeak is not simply a set of LDA topics: revolutionary themes appear in Marx, Lenin, Stalin, and Mao, for example, but their writing styles are recognizably different. Neither is it a matter of function words, since Maospeak exudes an affective strength and a clarity of purpose that cannot be reduced to

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 76–81
December 1-3, 2023 ©2023 Association for Computational Linguistics

Corpus	Number of Tokens (Words)	Vocabulary Size	Total Characters	Type-Token Ratio
Maospeak	1,684,294	55,113	2,890,605	0.0327
Contemporary	18,219,475	515,468	28,509,185	0.0283
Eileen Chang	994,025	70,484	1,532,465	0.0710
Mo Yan	2,546,989	131,317	3,978,079	0.0516

Table 1: Dataset Statistics

the most frequent grammatical particles alone. While thoroughly researched by scholars in humanistic disciplines (Ji, 2003; Link, 2013; Schoenhals, 2007) and vividly debated on public platforms (Sun, 2012; Link, 2012; Laughlin, 2012; Barmé, 2012), so far there has been very little computational engagement with the language of Mao Zedong (Huang and Shi, 2022). Maospeak poses an interesting challenge to modern stylometry and remains to this day a controversial issue, especially given the diverse opinions surrounding writers like Mo Yan (b. 1955), the 2012 Nobel Prize laureate who has been accused by critics of inheriting the Maoist style in his descriptions of war-time violence and brutality (Link, 2012; Sun, 2012).

2.1 Dataset

Our dataset (Table 1) consists of four corpora: the selected works of Mao Zedong, collected novels and short stories of Mo Yan and Eileen Chang (Zhang Ailing), and a larger compilation of 102 novels published by 62 writers active in the post-Mao era. In this study, we treat Mao Zedong's writings as a proxy for Maospeak, as it was chiefly through quotations from Mao that the discourse of class struggle and popular militarization spread across the PRC, thus shaping the everyday language. This influence was particularly evident during the Cultural Revolution (1966-1976), when inability to quote the *Little Red Book* could be taken as proof of reactionary politics (Ji, 2003, 151). We chose Eileen Chang (1920-1995) as a control writer because she spent most of her life outside mainland China, and her writings arguably lack communist influences; the mixture of contemporary Chinese writing serves as another control, offering a sample of modern literary Chinese. We preprocessed Mao Zedong's writings by removing footnotes and lines of text shorter than 50 characters to filter out titles, dates, and signatures frequently attached by editors to his letters and communiques. Since Chinese does not use spaces between words, all texts have been segmented with the `spaCy` parser for Chinese `zh_core_web_lg`.[1]

2.2 Perplexity

One way in which different literary styles can be compared is by evaluating the perplexity of their representative texts. A low perplexity indicates that the text is more predictable (Kilgarriff and Rose, 1998; Józefowicz et al., 2016). Auto-regressive language models such as GPT are especially useful in this regard: in the pre-training phase, the model iterates over a large amount of data and learns to predict the next token given a sequence of tokens; these learned predictions can be then used to calculate the "surprisingness" of the actual words encountered in a text.

The formula to calculate the average per-token perplexity for a corpus C consisting of M texts, each containing K words, is represented as follows:

$$\mathcal{P}(C) = \exp\left(-\frac{1}{T}\sum_{j=1}^{M}\sum_{i=1}^{K}\log p(w_{ji}|w_{j,1:i-1})\right) \tag{1}$$

In (1), $p(w_{ji}|w_{j,1:i-1})$ represents the probability of the i-th word in the j-th text of the corpus, given the preceding words in that text. The equation computes the average of log probabilities across all non-special tokens T. If the model assigns high probabilities to the actual words, the average log probability will be less negative, which, after negation and exponentiation, will lead to a lower perplexity. Conversely, lower probabilities for the actual words will result in a higher perplexity.

In this experiment, we used three publicly available Chinese GPT-2 models: **Wenzhong 2.0**,[2] **uer-gpt2**,[3] and **gpt2-base-chinese** from **CKIP Lab**.[4] Using multiple models allowed us to not only compare the results but also mitigate the impact of the pretraining data: some models might have "seen" Mo Yan's writings during pretraining, for example, which could lead to lower perplexity for Mo Yan's tokens. Given that GPT tokenizes Chinese by individual characters, we sampled 500-character sequences from all four classes. The sampling process involved random selection of 3,000 sequences from each class to ensure the unbiased representation of texts.

The results (Figure 1) show that Maospeak features

[1] *https://spacy.io/models/zh*

[2] **Wenzhong 2.0**, 3.5B parameters, pre-trained on the 300GB version of the Wudao Corpus which includes mostly internet-based content (Zhang et al., 2022).

[3] **Chinese GPT2-xlarge**, 1.5B parameters, pre-trained on the 14GB CLUECorpusSmall corpus which includes news, Wikipedia, and social media content (Zhao et al., 2023).

[4] **CKIP GPT2 base-chinese**, 102M parameters, pre-trained on traditional Chinese data including Wikipedia and the CNA (Central News Agency) corpus. We have used the Python package `OpenCC` for simplified-traditional conversion.

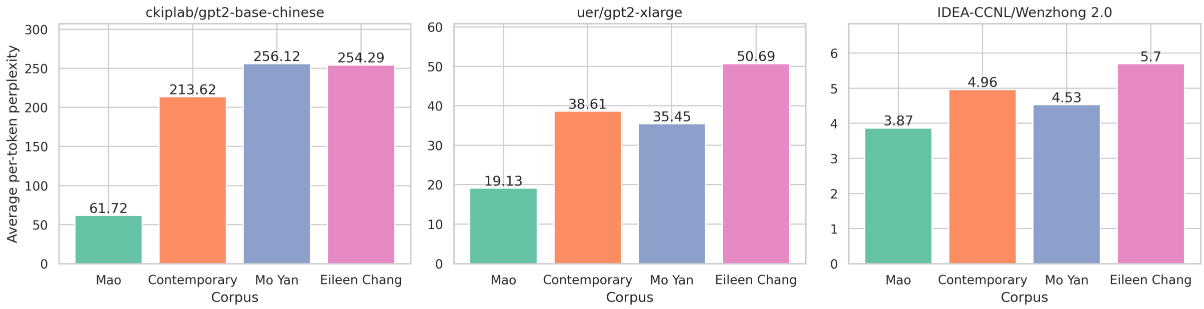

Figure 1: Average per-token perplexity across different GPT models.

a much lower perplexity than the other three classes, indicating a higher level of conformity with the language models' training data. By contrast, Eileen Chang's writings exhibit a very high per-token perplexity in all three models, reflecting more unexpected and creative word choices. The relatively high perplexity, and thus unpredictability, of Mo Yan's writing raises a question to consider for his critics, whereas the low perplexity of Maospeak hints at an important aspect of engineered languages (Ji, 2003), which promote the use of stock phrases and discourage creative usage of words (coining new metaphors, using rare vocabulary, unconventional syntax, etc). From this perspective, Maospeak resembles "machine text" rather than "human text," a distinction elaborated by Holtzman et al. (2019) in their work on neural text degeneration.

2.3 Systematicity

Another stylometric feature that differentiates texts from different authors is systematicity. Our hypothesis is that an author characterized by a high degree of systematicity would manifest a consistent overarching idea across all of their works. Essentially, each piece of a highly systematic writing can be viewed as a "microcosm" reflecting the broader semantic "macrocosm," even though individual texts may employ varied vocabulary.

One possible approach to measuring systematicity is to compute the average divergence between the overall ("global") vocabulary distribution across all texts produced by a given author and the ("local") vocabulary distribution in each of their specific writings. This methodology shares similarities with authorship attribution techniques such as z-scores of function words understood as an author-specific signal (Evert et al., 2017) which can be compared with a text-specific distribution through distance measures. However, our goal here is not to identify the real author among many possible ones, but to measure the particular author's thematic coherence.

The Kullback–Leibler (KL) divergence between the two probability distributions P and Q is given by the formula:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \cdot \log\left(\frac{P(i)}{Q(i)}\right) \qquad (2)$$

For discrete probability distributions P and Q, the KL divergence quantifies the amount of information lost when Q is used to approximate P. In other words, if an event is likely under P but unlikely under Q, the term $\log\left(\frac{P(i)}{Q(i)}\right)$ is large, contributing significantly to the overall KL divergence.

To calculate the KL divergence for our datasets, we randomly sampled 3,000 segments from each of the four classes, with each segment containing 500 words. Here, P represents the local distribution of words in a specific segment, while Q is the global distribution of words across the entire corpus corresponding to the given class, including the unsampled fragments. For each class, the global and local vocabularies are uniquely determined within that class, i.e., there is no one shared vocabulary built at the outset.

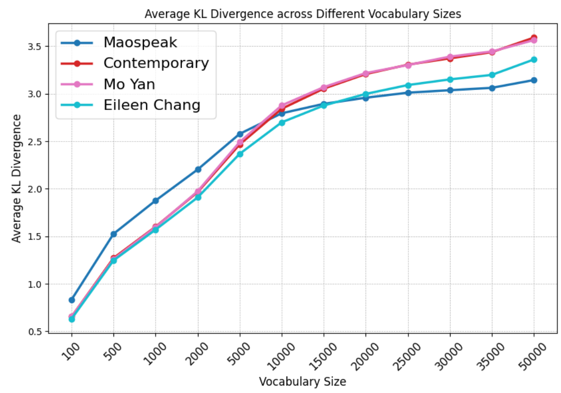

Figure 2: Average KL divergence across different vocabulary sizes; scale modified.

A few preliminary results encouraged us to conduct a series of tests and observe how the KL divergence changes as a function of vocabulary size. Surprisingly, Maospeak features the highest divergence for relatively small vocabulary sizes, which clearly separates it from the other three classes. This superiority diminishes, however, as we increase the vocabulary size (Figure 2). We confirmed the same results for segments of other lengths (100 and 1,000 words). A possible explanation of this behavior is as follows: when the average KL divergence continues to grow with the increasing vocabulary for the other three corpora (Mo Yan, Eileen

Chang, and the Contemporary corpus), it suggests that those texts feature a wide and diverse range of topics and themes. Each increase in vocabulary size continues to uncover more disparity between local and global distributions, which could potentially signify that these corpora are rich in specialized terminology or have a diverse set of topics or themes covered within them, and that less-frequent terms are distributed less uniformly across the 500-word segments.

By contrast, when the growth of divergence slows down with increasing vocabulary size, it may imply that the given corpus is more homogeneous and that most of the diversity or variability in word usage is captured at a smaller vocabulary size. Beyond a certain point, increasing the corpus-specific vocabulary size does not contribute significantly to revealing new disparities between local and global distributions. This suggests that the content of the Maospeak corpus is more focused and limited to a few main themes or topics, less-frequent terms being distributed more uniformly. At higher vocabulary sizes, Mao-style prose embodies the famous adage that "one sentence equals thousands of sentences" (一句顶一万句). The more we read Mao, the less we need to read Mao, since the amplitude of divergence is relatively small, whereas in the realm of literature every novel creates a new unexpected world. This last point holds as true for Mo Yan as for any other contemporary Chinese writer.

The above results suggest that it is not enough to compare the global-local divergence at a single vocabulary size. Eileen Chang, for example, exhibits the lowest divergence for middle-range vocabulary among all four classes, suggesting the relative coherence and internal similarity of her works, but her sentences and phrases become less alike once we take a larger vocabulary into account. In this sense, systematicity is an author-specific function of vocabulary size.

2.4 Characteristic words

Yet another method of measuring explainable differences in literary styles is to classify texts based on the presence or absence of characteristic words and expressions. TF-IDF, short for Term Frequency-Inverse Document Frequency, is a numerical statistic that reflects how important a word is to a document in a corpus. It is defined as follows:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D) \quad (3)$$

Where:

- t is a term (or word)
- d is a document containing the term
- D is the corpus or collection of documents

The Term Frequency (TF) is calculated as:

$$\text{TF}(t, d) = \frac{\text{Count of term } t \text{ in } d}{\text{Total terms in } d} \quad (4)$$

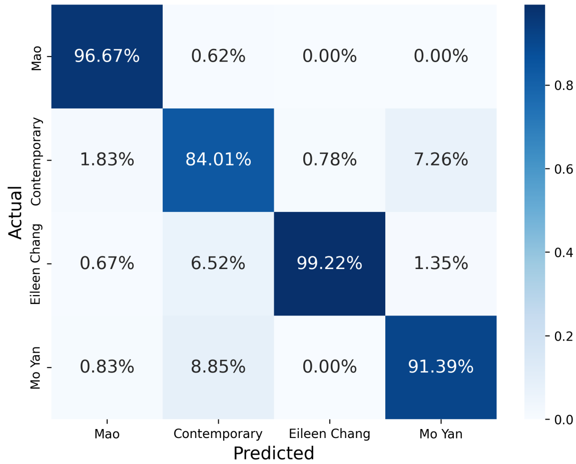

Figure 3: Confusion matrix for the Random Forest classifier with 300 TF-IDF features. Values have been normalized over the predicted conditions (columns).

The Inverse Document Frequency (IDF) is calculated as:

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total docs in } D}{\text{Docs with term } t}\right) \quad (5)$$

In other words, if a term t is frequent locally (4) and rare globally (5), its TF-IDF in the given document will be large. By utilizing TF-IDF, we emphasize words that are unique to a particular document while giving less weight to words that are common throughout the entire corpus. Training an explainable classifier, like a Decision Tree or Logistic Regression model, on such terms, enables the identification of features that most strongly indicate a particular style. In contrast to the two previous experiments, the features discovered in this way are relational and thus do not tell us anything about the particular literary style "as such," providing instead a way to distinguish different styles from each other.

The dataset in this experiment was built by dividing the four corpora into 500-word segments and then sampling 3,000 segments from each of them without replacement. In cases where fewer segments were available, we did not oversample. For example, given the smaller size of Eileen Chang's corpus, only 1,502 segments were obtained. All of the sampled fragments were put together and split into training (80%) and test sets (20%). We then trained a Random Forest classifier with 100 trees (estimators), each with a maximum depth of 15, the 300 words with the highest TF-IDF values obtained from the training data serving as our feature set. TF-IDF values were computed using the `TfidfVectorizer` from `scikit-learn`. We achieved 91.5% accuracy on the test set (2,092 examples), suggesting a relatively high degree of reliability in distinguishing different forms of writing, in particular those of Mao Zedong and Eileen Chang (Figure 3).

To gain deeper insights into which words are most indicative of a particular literary style, SHAP (SHapley Additive exPlanations) values have been computed post-training. SHAP values allow for the measurement of the impact (the average marginal contribution) that each feature (in this case, a word) has on the model's output (Mosca et al., 2022). For example, the computed SHAP values can reveal which words have the most influence in classifying a text as coming from the Mao Zedong corpus, essentially pointing out the vocabulary that distinguishes Maospeak from other styles.

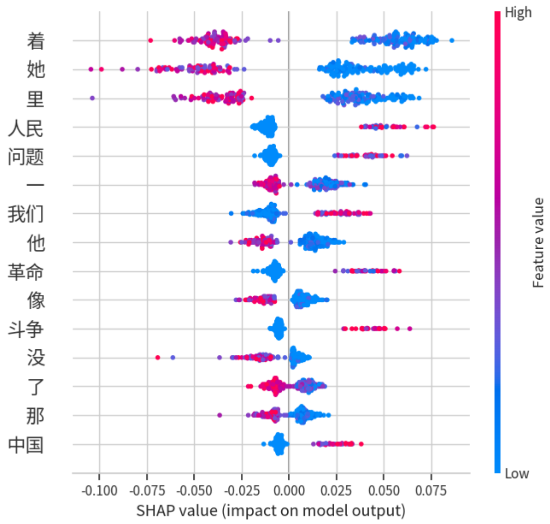

Figure 4: SHAP Values for class "Mao" in the Random Forest classifier, 200 test samples.

As shown in Figure 4, the computed SHAP values locate the main difference between Maospeak and the world of literature in the point-of-view markers: whereas literary texts are characterized by pronouns ("he" 他, "she" 她, "you" 你) and grammatical particles ("-ing" 着, "-ed" 了), which ground narratives in the actions and thoughts of individual characters, Mao-style prose "speaks" on behalf of the first-person plural "we" (我们) and gathers depersonalized, political terms such as "the People" (人民), "China" (中国), or "struggle" (斗争). Crucially, what the SHAP values demonstrate is that literary style is not only defined by the features present in a text but also by those that are absent, as shown by the blue dots which contributed (when absent, i.e., when bringing the TF value to zero or close to zero) to the model's final predictions. In this sense, it is the lack of point-of-view markers that characterizes Maospeak and the lack of political terms that characterizes contemporary prose. The role of absence within authorial signal is often overlooked by stylometric interpretations focusing solely on what is visible in the text, rather than what is not.

3 Conclusion

In this paper, we have analyzed three different aspects of Mao-style prose: perplexity, systematicity, and words with the highest TF-IDF values. The results of these experiments demonstrate some of the important features of Maospeak, an engineered language which reinforces the ideological tenets of Maoism through its formal characteristics. The conducted experiments also offer partial evidence to question the alleged Maoist influences on Mo Yan. While his violent literary style reflects China's revolutionary experience, it is hardly comparable to the redundant Party parlance.

While our analysis pertains chiefly to Maoism, we believe that our findings will be applicable also to more recent contexts, in China and beyond (Barmé, 2012). In particular, evaluations of next-token perplexity and KL divergence underscore the pivotal role of originality and subjectivity in language use. From this perspective, fiction reading and humanistic education become especially important. Reading widely and increasing one's exposure to various language data counters the influences of ideologies on our linguistically mediated perceptions of the world and increases the perplexity of our imaginations.

Limitations

Measurements of next-token perplexity are constrained by the availability of advanced hardware, the accessibility of large language models, the types of these models, and the amount and types of data that these models have been pre-trained on. In particular, tokenization proved a crucial consideration behind our choice of GPT models. In contrast to character-level tokenization used by Chinese versions of GPT, other tokenizers such as `SentencePiece`, used by generative models `CPM` (Zhang et al., 2020) and Chinese `LLaMA` (Cui et al., 2023), treats certain multi-character words (社会 "society" or 资本 "capital," e.g.) as single tokens, some of which are more prevalent in Mao's corpus compared to other corpora like that of Eileen Chang. In our tests, such discrepancies impacted measurements of perplexity, as the multi-character words are on average much less likely (i.e., they score lower probabilities) and thus increase the overall perplexity. Although the character-level tokenization of GPT models avoids this bias, treating each Chinese character individually and thereby providing a more uniform analysis across different writing styles and corpora, our choice of the pre-trained GPT models had a direct impact on the final results. Further analysis is needed to compare different models, tokenizers, and training data.

Acknowledgments

The author would like to thank the anonymous reviewers for their comments and suggestions. Special thanks are due to Aaron Gilkison and Heidi Huang for their feedback and support.

References

Geremie R. Barmé. 2012. New China Newspeak 新华文体. *China Heritage*. Online; accessed

09/25/2023.

J. F. Burrows. 1987. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2):61–70.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32:4–16.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Libo Huang and Xinyu Shi. 2022. A corpus-based investigation of the english translations of mao zedong' s speeches. *Frontiers in Psychology*, 13.

Fengyuan Ji. 2003. *Linguistic Engineering: Language and Politics in Mao's China*. University of Hawaii Press.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 46–52, Palacio de Exposiciones y Congresos, Granada, Spain. Association for Computational Linguistics.

Charles Laughlin. 2012. What Mo Yan's Detractors Get Wrong. *ChinaFile*. Online; accessed 09/25/2023.

Tuo Li. 1998. 汪曾祺与现代汉语写作——兼谈毛文体 [Wang Zengqi and Modern Chinese Writing: Also on the Style of Mao's Writings]. 花城 *[Huacheng]*.

Perry Link. 2012. Politics and the Chinese Language: What Mo Yan's Defenders Get Wrong. *Asia Society*. Online; accessed 09/25/2023.

Perry Link. 2013. *An Anatomy of Chinese: Rhythm, Metaphor, Politics*. Harvard University Press.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6).

Weihan Ou, Steven H. H. Ding, Yuan Tian, and Leo Song. 2023. Scs-gan: Learning functionality-agnostic stylometric representations for source code authorship verification. *IEEE Transactions on Software Engineering*, 49:1426–1442.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *CoRR*, abs/1609.06686.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michael Schoenhals. 2007. Demonising Discourse in Mao Zedong's China: People vs NonPeople. *Totalitarian Movements and Political Religions*, 8(3):465–482.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Anna Sun. 2012. The Diseased Language of Mo Yan. *The Kenyon Review*. Online; accessed 09/25/2023.

F. J. Tweedie, S. Singh, and D. I. Holmes. 1996. Neural network applications in stylometry: The "federalist papers". *Computers and the Humanities*, 30(1):1–10.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. Cpm: A large-scale generative chinese pre-trained language model.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*.