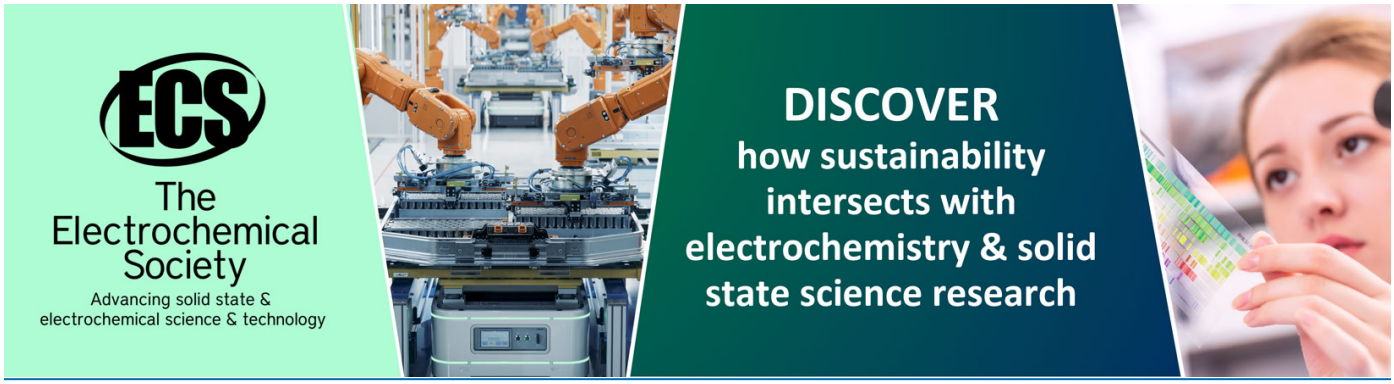
Comparing genres of Indonesian literature by utilizing big data



IOP Conf. Series: Materials Science and Engineering **434** (2018) 012277 doi:10.1088/1757-899X/434/1/012277

Comparing genres of Indonesian literature by utilizing big data

M Rokib[1*], S Anam[2], E Rahmawati[2], S Sodiq[1] and P Retnaningdyah[2]

[1]Universitas Negeri Surabaya, Indonesian Language and Literature Department, 60213 Surabaya, Indonesia
[2]Universitas Negeri Surabaya, English Language and Literature Department, 60213 Surabaya, Indonesia

*mohammadrokib@unesa.ac.id

Abstract. An academic study of literature generally requires researchers to understand literary genres, writers, periods, and even ideological clashes. A researcher is possibly able to master a wide range of literature dated back from the biblical period up to now in detail. In a certain case, comparing literary genre in a particular period provides a general understanding of literary trends. This paper offers a comparison of Indonesian literature genres by applying computational analysis in the context of Big Data utilization in Indonesian literature. Through the computational analysis, literary genres will be compared along with its popular authors following certain periods and trends. By transferring various literary works into the data, this analysis yields a diverse map of literary works that emerged in the Internet. The data reveals the relationship of literary characters that change over time. This paper argues that the use of computational analysis in the context of big data can be very helpful for scholars to know quickly and succinctly the Indonesian literature genres which is marked by certain categorization.

1. Introduction

For centuries, intuition and intelligence have played a central role in the study of literature. The emergence of big data that glimpsed both in humanities and natural sciences has offered a new approach which can be applied to traditional disciplines, such as literary studies. In the field of humanities, big data is also expected to update scientific disciplines which explore human thinking and cultural products. This new area of the scientific study often refers to Digital Humanities.

Is it possible if we analyze literary works through big data? Then, what is the importance of studying literature, for example, with big data? How does big data contribute to the study of literature, particularly Indonesian literature? These questions become a central point to be discussed in this preliminary discussion paper on challenges and opportunities of big data within literary studies.

The literary discipline is actually in accordance with the big data concept. Not just because the literary discipline has been very old and even the oldest in the field of humanities, but also because it has millions of works, writers, genres, and themes that already exist, were used and discussed extensively offline and online. It means that there are million or even billion data available in the literary disciplines. Besides, Big Data is essentially the huge amount of texts so that it can empower a new breed of humanities scholars [1].

In recent years, we have sought to look into the possibility and relevance of big data within literary studies. Big data is becoming a trend in various scientific disciplines, which can be applied to literary studies. This effort emerged and was first inspired when reading the work of Franco Moretti. His work describes wide range of literature dated back from the biblical period up to now in detail by showing it in the graphs, maps, and tress. This work not only re-questioned the concept of a very Eurocentric novel and why novels always stick to prose, but also open the eyes that in fact the computer can help to chew thousands of data of literary texts from century to century at a time. Moretti acknowledges that his research tends to be different and anti-mainstream because it works quantitatively to extract data on literary types, titles of literary works, writers, and so on with the treat of diagrams, maps, tables, and the like. To get to that result, the computational processing data aid has supported and facilitated the work [2]. While not explicitly mentioning the big data in his book, what Moretti did is essentially insightful for how big data works. An article written by Jennifer Schuessler alludes to the big data approach in literary studies with the assumption that Moretti refers that literary studies tend to analyze a limited choice of works at a particular time [3]. The selected work is just a sample. A literary critic needs to consider tens of thousands of other works at either the same or different time.

The challenge leads to an answer that literary scholars need the computer algorithmic assistance and digital data to map the history of literature over a long period of time. This is also what convinces a new criticism called computational criticism. In the study of literature, the way big data works will be different from the way other works that has been common in the approach of literary studies. We are familiar with various *isms* or theories in literary studies such as formalism, Freudianism, structuralism, post-colonialism, and postmodernism. The *isms* are the variety of isms that have always been used to interpret literature, politics or culture. In contrast to the isms, big data is not for the analysis of meaning of a particular symbol or symbol, but more deeply into what keywords are used by people from century to century. Big data is believed to have been able to serve readers, critics and literary lovers of billions of literary data through certain keywords with quantitative visualizations such as graphics and maps.

This approach emphasizes the networking aspect of literary discourse. The meaning of the network can be achieved by another approach, but with treats of big data mapping, certain meanings can be easily elaborated. It also does not close the possibility to use another approach if we want to explore a particular topic or case. One of the historians named Anthony Grafton, as quoted by Patricia Cohen, states that he strongly believes in quantification. He also admired how the field of digital humanities has done fantastic work. But he does not believe that quantification can do anything because humanitarian problems in the humanities are about meaning, of interpretation [4]. This is almost impossible by quantitative means.

In the similar context of literature and the use of data, one of the Indonesian literary experts published computational critic-based research. He utilized the Google search engine to measure the popularity of the writer Pramoedya Ananta Toer on the internet. With a quantitative approach enriched with descriptive approaches, he found that the novel *This Earth of Mankind* became very popular with five factors that support the perception of the novel [5].

The results of mapping using big data in literary studies are actually presenting a certain understanding, although it is not as deep as the way the traditional literary approach works. The most important point in this approach is not on the meaning of a particular work or topic in the work but how to understand the productive tidal productivity, the tendency of literary themes throughout history, genre tendencies over time, and the development of literary topics and genres over time.

Even though the works of Moretti and Graf show the practice of literary data mining in terms of literary studies through different method and computational analysis, however, the use of online statistical analysis together with data inside it has not received due attention. In the Indonesian literature context, to be specific, it has been neglected by many scholars. Thus, this paper attempts to provide a preliminary analysis of Indonesian literature genres comparatively in a long period using online computational analysis.

2. Methods

This research focuses on the comparative genre analysis of Indonesian Literature in a long period of time. Presenting the genres comparatively, the use of data in each period is compulsory. To gain the data before presenting the comparison of genres of Indonesian literature, computer-assisted analysis is used. This step is the practice of computational text analysis [6] [7] or computational criticism [8]. Here, we use data about the various genres in Indonesian literature which have been discussed and published by poets and scholars as they are documented in google books either in Indonesian or English.

There are a lot of books in google books related to genres in Indonesian literature since 1900s to 2000s. By putting semantic words of Indonesian genres, book titles, sub-titles, and related topics will be collected. To be specific, this research focuses on different genres and put genre categories together in an online application called google book 'ngram' viewer. This online application can organize the data, particularly Indonesian literary genres from a long period. There are statistical operations performed on the data that can present genres comparatively in a graphic model.

In the first step of research, various genres in Indonesian literature was identified. In this step, Sumarjo & Saini's classification of genre is employed. There are two domains, namely imaginative and non-imaginative genres in terms of modern Indonesian literature [9]. However, this research considers classic period or Malay literature. Some genres such as *pantun*, *gurindam*, *kisah*, and *hikayat* are excluded from the modern classification. To accommodate the genres of classical literature, this research put it in the list of Indonesian literary genres.

Secondly, this research put every single genre in a bar of google books ngram viewer. This program then generates every semantic word and identifies them from 1900s to 2000s. Every change or trend can be traced in each year within a graphic model. Because there are some genres, the graph showed genres comparatively as seen in the figure 1 as an example.

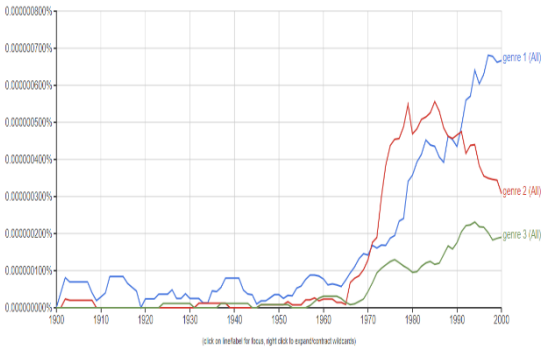


Figure 1. Ngram viewer model.

If we just simply click each genre, the program will redirect to the list of books semantically related to the topic or title of the books. Thus, researchers can analyze the trend as well as period of genres' popularity or perception. Detailed analysis of contents within each genre is not presented here because our research focus is only on the comparative genre analysis in the period of 1990s to 2000s.

3. Results

Following the Sumarjo & Saini's classification of Indonesian literary genre, there are some genres classified as imaginative and non-imaginative genres. In the non-imaginative genres, there are seven genres including essay, criticism, biography, autobiography, history, diary, and letters. In the imaginative genres, there are two categories, namely *puisi* or poetry and *prosa* or prose. *Puisi* includes epic, lyrics, and dramatic, while *prosa* include novel, short story, novelette, comedy, tragedy, melodrama, and tragedy-comedy.

Comparing classical genres of Indonesian literature, Sumarjo & Saini's classification neglects the existance of Malay literature genre which is popular in Indonesian society before Indonesia's

independence. In the *puisi* category, there are genres of *mantra, pantun, karmina, gurindam, seloka, syair,* and *talibu,* while in the *prosa* are *hikayat*, *kisah*, *dongeng*, and *cerita rakyat*. Those classical genres are popular before independence and even recent time though in different forms. There are no reason to abandon the classical genres in terms of literary history in Indonesia. Those genres should be included in the Indonesian literary genre as part of native's literary works. As can be seen in table 1, the whole Indonesian literary genres within the *puisi* and *prosa* category, either classical or modern genre, are twenty genres.

Table 1. Indonesian literary genres.

Category	Classical Genres	Modern Genre
Puisi (Poetry)	Mantra	Epik
	Pantun	Lirik
	Karmina	Dramatik
	Gurindam	
	Seloka	
	Syair	
	Talibu	
Prosa (Prose)	Hikayat	Novel
	Kisah	Cerpen
	Dongeng	Novelet
	Cerita Rakyat	Komedi
		Tragedi
		Melodrama

Of the twenty genres, we enter each category into the google books ngram viewer program. Each genre we enter together to derive a comparison between one genre and another in a certain category. The results of the comparative genre of poetry from 1900 to 2008 show an upward trend. Only one genre has increased from time to time, namely *mantra* genre. Compared to other genres, *mantra* is the most popular term followed by *pantun, epik,* and *syair*. However, after clicking the detailed information on the availability of data in the google book, the term *mantra* is not only mantra in Indonesian literature but also mantra in Indian and other terms. This finding made us exclude mantra from the comparative genre analysis. As can be seen in figure 2, the most popular is *pantun* followed by *epik* and *syair*.

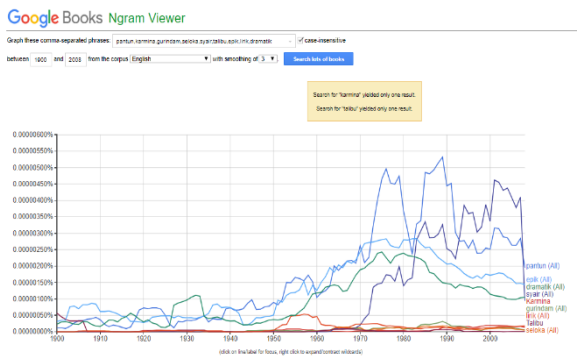


Figure 2. Comparative genre analysis of poetry category.

On the other hand, prose category showed that *hikayat* is the most prominent genre in the Indonesian literature in google book. Of course, there are error analysis when putting the term novel

and melodrama because both terms are too general so that the ngram program showed inaccurate results of Indonesian literary genres. Thus, we exclude the genre of novel and melodrama. As can be seen in figure 3, *hikayat* remains the most popular and the most used term in google book. It means that *hikayat* is the popular genre compared to others within prose category since 1900 to 2008 in the context of Indonesian literature.

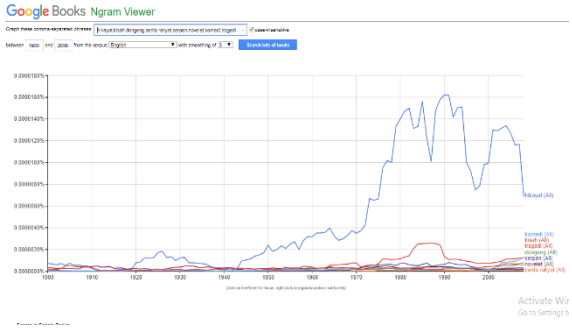


Figure 3. Comparative genre on prose category.

After comparing each category, comparing all genres either poetry or prose category is interesting. This comparison means that excluding three genres of novel, melodrama, and mantra is possible because it can destruct the result. Those terms are too general and did not reflect Indonesian literature genres only. The result showed as in figure 4 that the most popular genre is *hikayat* which has 0.0000116575% in 2007. This genre followed by *pantun* and *syair* in the same year. However, the most interesting result is the *pantun* and *syair* comparison which located *syair* as the most popular in 2007 which has 0.0000037797%, followed by *pantun* which has 0.0000026326%. In general, particularly in 1975, 1978, and 1988, *pantun* was more popular than *syair*.

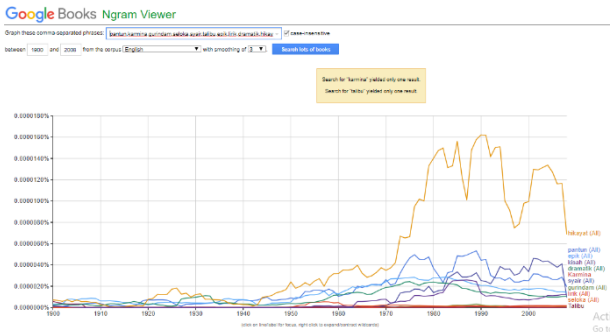


Figure 4. Comparative genre analysis of Indonesian literature.

4. Conclusion

In the present paper, the authors tried to compare Indonesian literary genres within prose and poetry category either classical or modern genres. By using google book ngram viewer program as a practice of computational text analysis in terms of Big Data, the most popular genre in Indonesian literature is *hikayat*. This genre is followed by *pantun* and *syair* as the other popular genres.

References

[1] Porritt G 2015 Data Mining in the Humanities and Social Sciences *Information Today* **32** 10 p 18
[2] Moretti F 2013 *Distant Reading* (London: Verso)
[3] Schuessler J 2017 *Reading by the Numbers: When Big Data Meets Literature* (The New York

Times) p C1
[4] Cohen P 2010 *Digital Keys for Unlocking the Humanities' Riches* (The New York Times) p C1
[5] Graf A 2007 Cyberpram: Pereptions of Pramoedya Ananta Toer on the Internet *Indonesia and the Malay World* **35** 103 p 293-312
[6] Y S and S K M 1986 *Apresiasi Kesusastraan* (Jakarta: Gramedia Pustaka)
[7] Huston N M 2014 Toward computational analysis of Victorian poetics *Victorian studies* **56** 3 p 498-510
[8] Boukhaled M A 2016 *On computational stylistics: mining literary text for the extraction of characterizing stylistic patterns* (Paris: Université Pierre et Marie Curie)
[9] Hackler R M dan Kirsten G 2016 *Distant reading, computational criticism, and social critique: an interview with Franco Moretti* (Foucoultblog, Zurich)