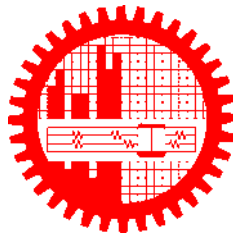


M.Sc. Engg. (CSE) Thesis

**A NOVEL WORD-TO-VEC GRAPH AND CHARACTER  
INTERACTION MODELS FOR LITERARY ANALYSIS:  
A CASE STUDY WITH BENGALI LITERATURE**

Submitted by  
Nafis Irtiza Tripto  
1017052002

Supervised by  
Dr. Mohammed Eunos Ali



Submitted to  
**Department of Computer Science and Engineering**  
**Bangladesh University of Engineering and Technology**  
Dhaka, Bangladesh

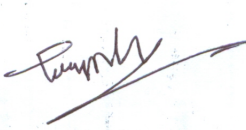
in partial fulfillment of the requirements for the degree of  
Master of Science in Computer Science and Engineering

March 2021

## Candidate's Declaration

I, do, hereby, certify that the work presented in this thesis, titled, "A NOVEL WORD-TO-VEC GRAPH AND CHARACTER INTERACTION MODELS FOR LITERARY ANALYSIS: A CASE STUDY WITH BENGALI LITERATURE", is the outcome of the investigation and research carried out by me under the supervision of Dr. Mohammed Eunos Ali, Professor, Department of CSE, BUET.

I also declare that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.




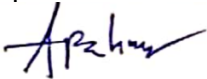

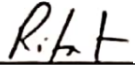
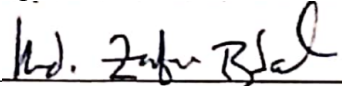
---

Nafis Irtiza Tripto

1017052002

The thesis titled “**A NOVEL WORD-TO-VEC GRAPH AND CHARACTER INTERACTION MODELS FOR LITERARY ANALYSIS: A CASE STUDY WITH BENGALI LITERATURE**”, submitted by Nafis Irtiza Tripto, Student ID 1017052002, Session October 2017, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfilment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents on March 18, 2021.

### **Board of Examiners**

1.   
\_\_\_\_\_
- Dr. Mohammed Eunus Ali  
Professor  
Department of CSE, BUET, Dhaka  
Chairman  
(Supervisor)
2.   
\_\_\_\_\_
- Dr. A.K.M. Ashikur Rahman  
Professor and Head  
Department of CSE, BUET, Dhaka  
Member  
(Ex-Officio)
3.   
\_\_\_\_\_
- Dr. Abu Sayed Md. Latiful Hoque  
Professor  
Department of CSE, BUET, Dhaka  
Member
4.   
\_\_\_\_\_
- Dr. Rifat Shahriyar  
Associate Professor  
Department of CSE, BUET, Dhaka  
Member
5.   
\_\_\_\_\_
- Dr. M. Zafar Iqbal  
Ex. Professor  
Department of CSE  
Shahjalal University of Science and Technology (SUST), Sylhet  
Member  
(External)

## **Acknowledgement**

First and foremost, I am thankful to God Almighty for the good health and well-being that he bestowed upon me.

Secondly, I thank my thesis supervisor, Dr. Mohammed Eunos Ali. Without his assistance and dedicated involvement in every step of the process from the very beginning, this work would have never been completed. I would like to thank him very much for his continuous help and support throughout the past years.

I wish to express our sincere thanks to Dr. A.K.M. Ashikur Rahman, Head of the Department, CSE, BUET, and previous Head, Dr. Md. Mostofa Akbar, for providing all the necessary facilities to complete our work. I take this opportunity to express gratitude to the department faculty members and my colleagues for their continuous encouragement and support. I especially thank Mohammad Mamun Or Rashid, Assistant Professor, Department of Bangla, for some of his suggestions in my early work.

Finally, I must thank my wife Mahjabin Nahar, lecturer of CSE, BUET for her continuous support and encouragement to conduct my thesis. I am also indebted to my parents for their unceasing encouragement, support, and attention.

Dhaka  
March 18, 2021

Nafis Irtiza Tripto  
1017052002

# Contents

<b>Candidate’s Declaration</b>	<b>i</b>
<b>Board of Examiners</b>	<b>ii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Algorithms</b>	<b>x</b>
<b>Abstract</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 State of the Art . . . . .	3
1.3 Research Objective . . . . .	4
1.4 Our Approach . . . . .	4
1.4.1 <i>Word2vec Graph</i> and Stylometry Tasks . . . . .	4
1.4.2 Character Interaction Graph for Interrelationship Analysis . . . . .	6
1.5 Contribution . . . . .	8
1.6 Organization . . . . .	9
<b>2 Related Works</b>	<b>10</b>
2.1 Solving Stylometry Tasks . . . . .	10
2.1.1 Author Attribution . . . . .	10
2.1.2 Genre detection . . . . .	11
2.1.3 Writing style change for authors . . . . .	11
2.1.4 Stylometry features . . . . .	12
2.2 Character Interaction Graph and Application . . . . .	13
2.2.1 Character Interaction Network . . . . .	13
2.2.2 Social Interaction Analysis from Character Network Graph . . . . .	14

2.3	Bengali Literature Analysis . . . . .	14
2.3.1	Stylometry Tasks in Bengali Literature . . . . .	14
2.3.2	Critical Analysis in Bengali Literature . . . . .	15
2.3.3	Character Network Analysis in Bengali Fiction . . . . .	15
<b>3</b>	<b><i>Word2vec graph Model for Stylometry Tasks and Literary Analysis</i></b>	<b>17</b>
3.1	The <i>Word2vec Graph</i> Model . . . . .	17
3.1.1	The <i>Word2vec graph</i> creation . . . . .	17
3.1.2	Features extraction from <i>Word2vec graph</i> : . . . . .	19
3.1.3	Clustering with <i>Word2vec graph</i> features: . . . . .	19
3.2	Experimental Evaluation . . . . .	20
3.2.1	Corpus Creation . . . . .	21
3.2.2	Baseline Methods . . . . .	23
3.3	Results and Discussion . . . . .	25
3.3.1	Result on Bengali Literature Corpus . . . . .	25
3.3.2	Result on Subset of Project Gutenberg Corpus . . . . .	30
3.3.3	Result Analysis on Newspaper Editorial Corpus . . . . .	31
3.3.4	Feature Selection for Various Feature set . . . . .	32
3.3.5	Discussion and Limitations . . . . .	36
<b>4</b>	<b>Understanding Social Structures from Character Interaction Graph</b>	<b>38</b>
4.1	Methodology and Experiments . . . . .	38
4.1.1	Character Interaction Graph Generation . . . . .	39
4.1.2	Graph Features Extraction . . . . .	43
4.1.3	Datasets . . . . .	45
4.2	Results and Findings . . . . .	45
4.2.1	Age and Gender Distribution . . . . .	46
4.2.2	Protagonist Characteristics . . . . .	49
4.2.3	Influence of Family . . . . .	50
4.2.4	Variation in Graph Structures . . . . .	51
4.3	Discussion . . . . .	52
4.3.1	Influence of Real-life Events in Story Character Arc . . . . .	52
4.3.2	Influence of Different Age and Gender Group . . . . .	55
4.3.3	Interpretation of Character Interaction Graph from Context and Genre . . . . .	56
4.3.4	Limitations . . . . .	57
<b>5</b>	<b>Conclusion</b>	<b>60</b>
	<b>References</b>	<b>62</b>

<b>A</b>	<b>Census Report</b>	<b>72</b>
A.1	Census Report in Contemporary times . . . . .	72

# List of Figures

1.1	<i>Word2vec graph</i> on three writings of Humayun Ahmed. The red, green and blue color represent <i>core</i> , <i>multiple</i> , <i>boundary</i> nodes/edges respectively (discussed in Chapter 3). The <i>core</i> and <i>multiple</i> nodes are densely connected for short story. However, <i>core</i> edges are absent in the <i>Word2vec graph</i> of both novel and long historical novel . . . . .	5
1.2	Character interaction graph on two novels of Rabindranath Tagore. The bigger the node or thicker the edge is, indicate more weight to the corresponding character or relation in the story. . . . .	7
3.1	Overall architecture of <i>Word2vec graph</i> and <i>Word2vec graph</i> +Graph words approach . . . . .	18
3.2	k-means clustering visualization for all authors using TF-IDF feature set <i>with-stopwords</i> version. Every point denotes a sample story. Each color represents a cluster and each marker represents the writing of an author. . . . .	27
3.3	k-means clustering visualization for all authors using TF-IDF feature set <i>without-stopwords</i> version. Every point denotes a sample story. Each color represents a cluster and each marker represents the writing of an author. . . . .	28
3.4	K-means clustering visualization for all authors using TF-IDF feature set <i>without-stopwords</i> . . . . .	32
3.5	Average value (with standard deviation marked) of some features in <i>Word2vec graph</i> for novels and short stories of various authors. . . . .	33
3.6	Average value (with standard deviation marked) of some features in <i>Word2vec graph</i> for different domains. . . . .	34
3.7	Distribution of sentence length in various writings. . . . .	35
3.8	Distribution of various stylometry features. . . . .	35
3.9	Change of different stylometry feature over writing years. . . . .	36



4.1	Interaction of different characters in a sample story chapter. Each sentence is denoted as a cell. A filled cell indicates that specific character is appeared in this sentence. Segement 1 and 2 of Character 3 are considered seperate segments because their distance is greater than $\delta_A$ . Similarly, segment 2 of character 1 and segment 3 of character 2 do not belong to the same plot since their distance is greater than $\delta_B$ . . . . .	40
4.2	Generation of story graph for novel দেবী চৌধুরানী ( <i>Devi Chowdhura,ni 1884</i> ) by Bankim Chandra from corresponding chapter graphs. The final story graph include all nodes and edges that are present at any of the chapter-wise graphs. The node and edge weights in final are computed from the weighted average of these graphs. Blue, green and red edges indicate neutral, positive and negative sentiment respectively. . . . .	43
4.3	Distribution of different age group over time for different authors. . . . .	48
4.4	Weight associated with family members for different authors over time. . . . .	51
4.5	Graph topological properties for different genres. . . . .	52
4.6	Character interaction graph for two novels of Bankim. Protagonist (Index 0) of both stories are widow. . . . .	53
4.7	Character interaction graph for two political novels of Rabindranath and Sarat influenced by nationalist movement. . . . .	54

# List of Tables

3.1	Extracted features from <i>Word2vec graph</i> . . . . .	20
3.2	Different authors with their characteristics and book counts on various genres in Bengali literature corpus. . . . .	21
3.3	Article counts and characteristics of different newspaper editorials . . . . .	22
3.4	Overview of our Gutenberg corpus . . . . .	23
3.5	Extracted stylometry features . . . . .	25
3.6	Weighted F1 score for various feature set in author identification task. . . . .	26
3.7	Weighted F1 score for various feature set in genre detection task . . . . .	28
3.8	Performance of various approach in clustering wrt timing of writing . . . . .	29
3.9	Result on author attribution in Project Gutenberg Corpus . . . . .	30
3.10	Result on genre detection in Project Gutenberg Corpus . . . . .	31
3.11	Performance on Newspaper corpus . . . . .	32
4.1	Overview of dataset for character interaction . . . . .	45
4.2	Age & Gender Proportion and Aggregate Weight for Each Group in Different Authors. All Values are Indicated in Normalised Form. . . . .	46
4.3	Average Degree Count of Different Age and Gender Group. . . . .	47
4.4	Age and Gender Wise Combined Distribution for All Authors. All Values Are Indicated in Percentage (%). . . . .	47
4.5	Age-wise edge distribution in different authors . . . . .	48
4.6	Male and Female Protagonist Characteristics for Authors . . . . .	50
4.7	Average Count of Nodes, Edges and Graph Density for Different Authors . . . . .	51
4.8	Year-wise Story, Genre and Protagonist Information for All Authors . . . . .	58
4.9	Political and social events in contemporary time of late nineteenth and early twentieth century [1,2] . . . . .	59
A.1	Census Report in Contemporary Time of late Nineteenth and Early Twentieth Century in Bengal [3]. All Values Are Indicated as Percentage (%) Form. . . . .	72

# List of Algorithms

# Abstract

Literature, as an imitation of human behavior, portrays the picture of society. Literary analysis offers a meaningful analysis of the literature by involving critical thinking from multiple perspectives. Analyzing the writing styles of authors and articles is a key to supporting various stylometry analysis tasks such as author attribution, genre identification, etc. Over the years, rich sets of features that include stylometry, bag-of-words, n-grams have been widely used to perform such literary analysis. However, the effectiveness of these features largely depends on the linguistic aspects of a particular language and the characteristics of the datasets. Techniques based on these feature sets cannot give desired results across domains. Consequently, social structures and real-world incidents often impact contemporary literary fiction. However, existing researches in literary fiction analysis explain these phenomena in a mostly non-technical perspective through the critical analysis of stories. Character networks (or graphs), in this scenario, can be particularly suitable for information retrieval from fiction to address various high-level problems.

In this study, we perform literary analysis from both perspectives by solving stylometry tasks as well as incorporating character networks. We are the first to utilize character interaction graphs to answer a wide range of social questions regarding the influence of contemporary society on literary fiction. Our study involves constructing character interaction graphs from fiction, extracting graph features, and exploiting these features to resolve these queries. Experimental evaluation of influential Bengali fiction over more than half a century demonstrates that character interaction graph can be highly effective in certain types of assessments and information retrieval from literary fiction. We also propose a novel word2vec graph based modeling of a story that can rightly capture both the context and the structure of the story. By using these word2vec graph based features, we develop a classification technique to perform several stylometry tasks: author attribution, genre detection, stylochrography. Our detailed experimental study with a comprehensive set of literary writings from famous authors of Bengali literature shows the effectiveness of this method over traditional feature based approaches.

# Chapter 1

## Introduction

As a timeless piece of entertainment, literature has an enduring effect on human lives. Literary fiction, particularly has been a significant part of human culture and often mirrors society and societal values [4]. Literature analysis illustrates the essence and greater intent of successful arts, whether to evoke empathy, encourage society, or simply to entertain. Therefore, literary analysis has received significant attention from both academia and industry over the years. Theoretically, literary analysis is the study, evaluation, and interpretation of literature — examining all parts of a story, such as word usage, characters, style, setting, tone. Literary analysis can be performed in various ways, such as identifying writing style from stories, which is commonly referred to as stylometry, and analyzing characters of stories from their interaction. These approaches can further encompass various literary analysis tasks that have an impeccable impact on various practical applications in natural language processing (NLP). However, for a low resource language like Bengali, such analysis from a computational perspective is not present. The literary analysis itself is a massive domain, and our research is limited to the approaches which are quantitative from different perspectives (text and characters).

In this study, we aim to perform literary analysis from both perspectives. We try to improve various stylometry tasks with a novel feature set and analyze the writings based on stylometry in Bengali literature. Also, we attempt to discover unique character arcs and find key insights into the social structure from Bengali literature fiction.

### 1.1 Motivation

The emergence of data analytic and machine learning techniques has enabled us to answer a variety of stylometry and social science questions by analyzing literary fictions. Therefore, different stylometry tasks have been popular in identifying the writing styles, techniques of an author over the last few decades. Examples include author attribution: identifying the author of particular texts [5, 6], verifying whether two documents are written by the same author [7, 8],

writing style change of an author [9–11]. These works have significant applications in real life, such as likability prediction of books [12, 13], literary book recommendations [14], literature analysis [15, 16], plagiarism detection [17], and many more. Exploration of writing style over time and writing resemblance detection among multiple authors/genres could provide much insight into the quantitative assessment for the literature of a particular language. Moreover, literary fiction present an elaboration of social norms [18], representation of the real-world [19], and assessment of social strategies [20] apart from their creative and recreational aspects,. The context, characters and storylines of contemporary literature are often determined by these factors. These analyses often require specific knowledge of that language and historicity and such assessment are not yet present in the Bengali literature.

Bengali, also known by its endonym Bangla, is an Indo-Aryan language with approximately 228 million native speakers and another 37 million as second language speakers [21, 22]. Bengali literature, has a millennium-old literary history and, is one of the most prolific and diversified literary traditions in Asia. Modern Bengali literature was developed during the 19th century and evolved significantly over time [23]. Bengali literature has produced many notable talents who flourished the language with their unique writing and creativity in diverse forms. However, there has been rarely any quantitative assessment in Bengali literature to specify author resemblance based on genre or topic, change of writing style over a period or discover exclusive stylometry features that actually represent their writing. Moreover, contemporary social structure and real-life incidents often influence the social representation depicted in literary fiction . For example, different historical events or the role of different gender and age group of characters can indicate the writer’s perspective on contemporary society and corresponding social structure [24–26]. However, these have not been yet investigated from a computational perspective. In fact, lack of adequate resources and corpus poses a great challenge in the computational evaluation of Bengali literature.

Therefore, literary analysis in Bengali literature can answer a wide array of questions regarding the effects of contemporary society, an assessment of characters, and social structures portrayed in fiction. Also, solving various stylometry tasks in Bengali literature can provide valuable insights into the writings of different authors, change of their writing styles, and effects of different stylometry features in their writings. Moreover, the success of different traditional feature sets in these various stylometry tasks heavily relies on the dataset size, characteristics, and language. So, a feature set independent of domain or language could facilitate different stylometry tasks across various territories of NLP. Therefore, it provides us a unique opportunity to solve different stylometry problems with a newly developed feature set as well as examine the writing styles, and discover character interaction from stories in Bengali literature.

## 1.2 State of the Art

Historically, literary analysis has been studied using a vast array of methodologies, ranging from qualitative treatments to the use of complex statistical techniques or data science. Therefore, stylometry research has yielded several methods and tools over the last years to handle a variety of challenging tasks.

**Research in stylometry tasks:** Most of the stylometry tasks can be generally considered as a text classification problem that assigns a label, i.e., author/genre/chronological timeline, to a text of various forms such as story, article, or writing. To classify a text in the literature and other related domains, a wide range of feature sets, such as bag-of-words, word n-grams, lexical, syntactic, semantic, structural attributes, and sequential modeling have been used [5, 27–31]. However, utilizing these distinct sets of features for classification has some key shortcomings, in particular for literary analysis of a specific language. These are: (i) A literary writing itself is not well structured like Wikipedia/newspaper article or online reviews. Instead of providing concise and straightforward information, a literary writing conveys intricate themes throughout a very long narrative [32]. Therefore, features that depend on the organizational structure of the document are not suitable for this domain. (ii) Many syntactic, semantic features are language-specific, and the extraction of these features from text requires specific procedures and understanding for that particular language. Also the efficiency of these features often depends on the characteristics of the datasets [33], and thus a feature that works for a specific language may not be appropriate for the literature domain of many other languages. (iii) Finally, similar types (e.g. genre) literary works of different authors could share common attributes and features. In such a case, classification algorithms would often lose information about the mutual relationship among these works by recognizing them as separate classes, which is essential in many literary analysis.

**Research in Bengali literature analysis:** Most of the present studies in Bengali literature analysis focus entirely on the author attribution part [34–37]. Furthermore, there has not been any quantitative assessment in Bengali literature to evaluate the influence of contemporary social structure and real-life incidents on the social representation depicted in literary fiction. Some studies address the role of women [2, 38, 39], the influence of nationalist movements at various times [40], particular views on religion [41, 42], and other subjects in the writings of various authors in Bengali literature. However, all of these existing works discuss these issues based on the manual and human expert involving non-technical perspective, which mostly consider the plot, description of characters, or critical analysis of the story to assert their hypothesis. They would often miss crucial information from a long story, unable to verify thoroughly how the author portrays his point of view through the plots & characters. These assessments also require

going through the whole story, and therefore they are limited to only a few stories by any specific author.

## 1.3 Research Objective

The primary intention of our thesis is to perform literary analysis on Bengali literature. More specifically, we aim to achieve the following two objectives in our research: (i) address different stylometry tasks, such as author attribution, genre detection, stylochronometry, and analyze the writing styles of authors from the story text (ii) comprehend the social structures from contemporary literary fiction in Bengali literature.

First, we intend to identify novel features for language or domain agnostic literary analysis. For this purpose, we aim to explore various stylometry problems in Bengali literature, such as author attribution, genre detection, stylochronometry with our newly proposed feature set and evaluate its performance across domains and languages. Also, it is our objective to present an analysis of the stylometry developments of authors and identify which features effectively contribute to finding distinction among various writings in Bengali literature.

Secondly, we aim to assess Bengali literary fiction from a character interaction perspective. For this purpose, we intend to identify the influence of social structure and contemporary events in fiction and answer a wide range of social questions regarding the influence of contemporary society on literary fiction from a data-driven perspective. It would enable us to make an interesting connection between real-world social structures & events and literary fiction.

## 1.4 Our Approach

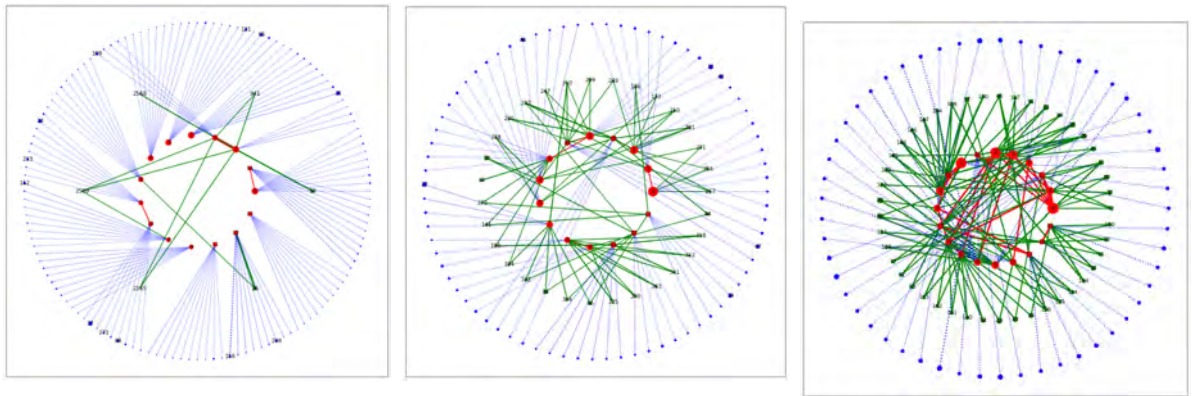
To achieve our research objectives, we conduct this study from two distinct viewpoints. Therefore, the approach in our dissertation consists of employing a novel feature set to solve various stylometry tasks. By using these features, we develop a classification technique to perform author attribution, genre detection, and stylochronometry tasks. Secondly, we utilize the character interaction model to answer various research questions in the context of Bengali literature. Our study involves constructing character interaction graphs from fiction, extracting graph features, and exploiting these features to resolve various queries.

### 1.4.1 *Word2vec Graph* and Stylometry Tasks

To overcome the previously mentioned limitations of existing feature sets in the literary analysis, we present a novel word-embedding graph (namely *Word2vec graph*) that can capture the underlying structure of a document. Each node of the *Word2vec graph* represents a word in the



document and an weighted edge between two words denote their similarities in the embedding space. The intuition of modeling the document as a *Word2vec graph* is that this graph structure can represent both the context of the document (from words associated with nodes) and the writing style of the author (from word usages and co-occurrences among words). Figure 1.2 shows a nice visualization of the *Word2vec graph* models of three different types of novels of a famous Bengali writer. From this example, we can see the clear structural difference among three types of writing of the same author. We are the first to identify this structural phenomena in the literature, which can play a key distinguishing feature in the literary analysis in any language.



(a) Historical novel মধ্যাহ্ন হান্ন্যো (Modhannyo) (b) Novel নন্দিত নরকে (Nondito Noroke) (c) Short story এলেবেলে (Elebele)

Figure 1.1: *Word2vec graph* on three writings of Humayun Ahmed. The red, green and blue color represent *core*, *multiple*, *boundary* nodes/edges respectively (discussed in Chapter 3). The *core* and *multiple* nodes are densely connected for short story. However, *core* edges are absent in the *Word2vec graph* of both novel and long historical novel

We extract a set of features that include node/edge weight, index, degree, neighbors etc., to capture the *Word2vec graph* structure of a document. These representative sets of features of a *Word2vec graph* enable us to perform unsupervised clustering based on these features and associates words, which play key role in different literary analysis. These features can reflect the overall narrative of the story through the use of words and the interaction between words. *Word2vec graph* creation and feature extraction can be performed for any document in any domain, regardless of the size of the document. Another major advantage of using *Word2vec graph* is that it can be applied across multi-lingual documents as the structure does not depend on the specific language.

We utilize *Word2vec graph* feature set in three major literary analysis tasks: author attribution, genre detection, and stylochometry. We evaluate the efficiency of the *Word2vec graph* feature set with both bag of words (unigram) feature set having Term Frequency Document Frequency (Tf-Idf) score, and feature set with various stylometry attributes (lexical, syntactic, sentiment, etc.). Both of them have been effective in these text classification tasks irrespective of domains [43]. In this paper, we focus on the literary analysis of eight prominent authors

(Rabindranath Tagore, Bankim Chandra, Sunil Gangopadhyay, Humayun Ahmed, etc.)<sup>1</sup> of Bengali literature from different periods. We also explore different feature sets to present an analysis of the stylometry developments of authors having a prolonged career and identify which features effectively contribute in finding distinction among various writings. To show the efficacy of the *Word2vec graph* feature set irrespective of domain/dataset characteristics, we also experiment with the editorial writings of several Bengali newspapers and a subset of Project Gutenberg corpus (English literature fiction). Therefore, we believe that the Word2vec graph could be an efficient approach to represent any text document, and corresponding words with the graph can be an alternative than utilizing all/most frequent words of the document in any NLP tasks.

### 1.4.2 Character Interaction Graph for Interrelationship Analysis

In this study, we take a computational approach that utilize the strength of data analytic approach to identify the impact of social structure, contemporary events in literary fiction. Modern approaches in narrative analysis focus on characters and their interactions [44]. A character's persona is constructed on how it interacts with other characters in different settings throughout the story [45]. Character interaction graph can portray these interactions as a node-link diagram and can be employed to assess different literary theories, identify the level of realism, and depict the social structure [46].

A character interaction graph or character network is a graph extracted from the story's narrative, in which vertices symbolize characters and corresponding edges represent interactions between them. Different attributes of nodes and edges indicate the characteristics of associated characters and relations. Figure 1.2 shows an intuitive visualization of the character interaction graph models of two novels of a famous Bengali writer. From this example, we can see the clear structural difference between these novels. The first one, being a romantic novel, indicate a strong relationship between the central male and female character. The other novel is a political one and contains more characters and interactions. The graph density and the number of characters with significant weight are also higher than the first novel. As seen in this toy example, the character interaction graph can often reflect the plot of fiction and we can more exploit more explicit and implicit features of the character interaction graph for further explorative analysis of the literary fiction.

In our research, we identify whether character interaction in fiction can depict the real-world social structure and perspective of the author. In particular, we aim to answer the following research questions in the context of Bengali literature.

---

<sup>1</sup>[https://wikipedia.org/wiki/Rabindranath\\_Tagore](https://wikipedia.org/wiki/Rabindranath_Tagore), [https://wikipedia.org/wiki/Bankim\\_Chandra\\_Chatterjee](https://wikipedia.org/wiki/Bankim_Chandra_Chatterjee), [https://wikipedia.org/wiki/Sunil\\_Gangopadhyay](https://wikipedia.org/wiki/Sunil_Gangopadhyay), [https://wikipedia.org/wiki/Humayun\\_Ahmed](https://wikipedia.org/wiki/Humayun_Ahmed)

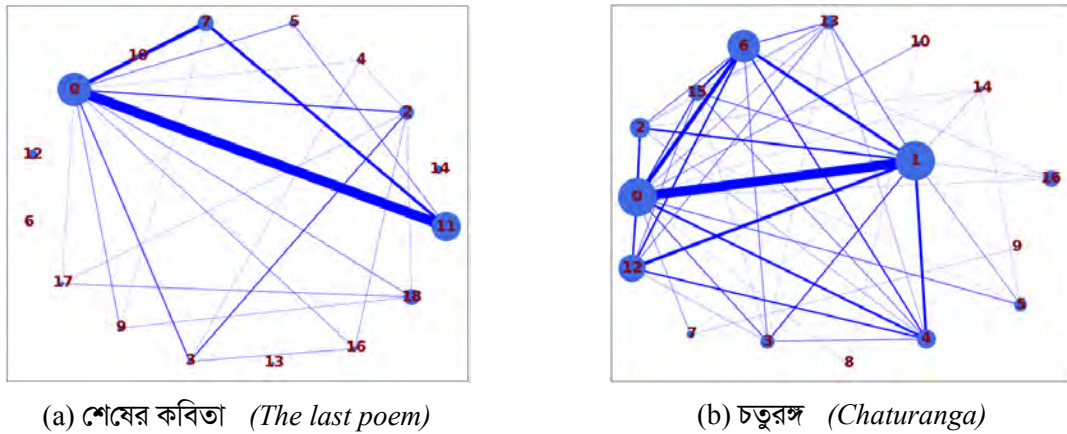


Figure 1.2: Character interaction graph on two novels of Rabindranath Tagore. The bigger the node or thicker the edge is, indicate more weight to the corresponding character or relation in the story.

- RQ 1: Have historical events influenced the story’s character arc and the dominance of characters in Bengali literature?
- RQ 2: Can we explain the influence of different age and gender groups in the Bengali society from novels of contemporary time?
- RQ 3: Can the presence and interaction of characters be interpreted by the story’s context and genre?

To answer all the the above questions, we rely on the novels of the three most prominent writers at the beginning of modern Bengali literature (Rabindranath Tagore, Bankim Chandra, Sarat Chandra Chattopadhyay) whose combined literary career span more than a half-century (1865-1935). We also consider several novels of two recent eminent writers of Bengali literature: Sunil Gangopadhyay and Humayun Ahmed to make a comparison with the modern period. To analyze the literary fiction, we first construct character interaction graphs of these novels from the character co-occurrence in the story narrative. Then we compute weight, sentiment score, and other attributes of nodes and edges from the story text as well as assign age, gender, role, and other information to different characters. To interpret the character interaction, we extract different node, edge, and graph features. Finally, we explore these features in chronological order for each author and evaluate whether real-world phenomena or social structures influence the character interaction in stories from different perspective.

Results of our study demonstrate that historical events like widow remarriage law in Hindu society (1872) and nationalist movements, such as the partition of Bengal (1906), non-cooperation movement by Gandhi (1920) influenced the character interactions and features in contemporary writings. Moreover, our results indicate that even if the presence of female characters was less than the male like prior researches in popular media [25, 47, 48], they hold similar or more weight

than male characters in that period. Also, the influence of older age groups decreases for authors who experienced various nationalist movements.

## 1.5 Contribution

The notable contributions of our study in both research directions are summarized as follows.

### **Solve stylometry tasks and analyze writing style:**

- We develop a novel *Word2vec graph* model to represent any document and utilize this graph structure and word features to solve various stylometry tasks. More specifically, we also solve author genre detection and stylochronometry problems in Bengali literature. We compare our proposed *Word2vec graph* based approach with both bag of words (unigram) and various stylometry feature sets, where our approach consistently outperforms them.
- We also identify the most prominent features from *Word2vec graph* and other two baselines that mostly contribute to various stylometry tasks and discuss the stylometry development of various Bengali authors.
- We show the efficacy of the *Word2vec graph* feature in a different domain by presenting separate case studies with the editorial writings of several Bengali newspapers and subset of Project Gutenberg corpus. Our feature set shows superior performance than both stylometry and unigram feature sets.

### **Identify social structures from literary fiction:**

- To the very best of our knowledge, we are the first to utilize character interaction graphs from literary fiction to answer a wide range of social questions regarding the influence of contemporary society on Bengali literature.
- We explore the effect of real-life events in story character arc, identify the influence of different age & gender group, and validate character interaction from story context & genre.
- By delivering visualization and quantitative assessment of influential fiction, our study can facilitate modern day researchers to perform critical literary analysis. Young writers can also gain more insight into how contemporary social structures can be portrayed in famous novels or how characters interact and thus enhance their writings.

## 1.6 Organization

The rest of our dissertation is organized as follows. We present an overview of relevant studies in Chapter 2. We discuss Word2vec graph model for solving stylometry tasks and perform various literary analysis in Bengali literature in Chapter 3. In Chapter 4, we utilize character interaction graphs to answer a wide range of social questions regarding the influence of contemporary society on literary fiction. Finally, we conclude the research in Chapter 5 and discuss our future works.

# Chapter 2

## Related Works

In this chapter, we discuss related works of our dissertation. First, we provide a brief analysis of some stylometry analysis related tasks and features. We also discuss character interaction graphs and their utilization in the social interaction analysis. Finally, we also review the existing works related to stylometry that have been done so far in the Bengali language, explore existing Bengali literature critical analysis and identify the scope of improvement.

### 2.1 Solving Stylometry Tasks

The analysis of authorial style, termed stylometry, assumes that style is quantifiably measurable for the evaluation of distinctive qualities. We review three prominent stylometry tasks that we aim to solve utilizing a novel feature set. Then, we briefly discuss different features that have prominent in different stylometry tasks.

#### 2.1.1 Author Attribution

Author attribution has been extensively studied over the last few decades, with a wide range of features and classifiers. Kjell et al. [49] first utilized the character bigrams and trigrams for authorship analysis. Later, the works in [5, 28] used character n-grams of different sizes and reported an accuracy of up to 72%. Bag-of-words approaches have also been successful for authorship attribution [50]. Various stylometry related features (lexical, syntactic, character, semantic, application-specific) are widely used in authorship attribution for different datasets [31, 51]. Guthrie [52] explored various features used for authorship attribution including commonly used stylistic features and several others intended to capture the emotional tone of the text.

However, effectiveness of these features relies on the characteristics of the datasets. Sari et al. [33] explored how different types of features affect authorship attribution accuracy under

varying conditions. Their study suggests that content-based features tend to be suitable for datasets with high topical diversity and datasets with less topical variance, benefit more from style-based features. Therefore, we consider both word unigram feature set to capture the content and stylometry features to find the stylistic signature of authors in both datasets and utilize them as baseline methods.

Different machine learning classifiers have been popular in the author attribution task, such as Decision tree [53], Support Vector Machine [54], Convolution Neural Network [30]. However, as previously discussed, we employ unsupervised simple K-means clustering to explore stylistic similarity among various authors using these feature sets. Clustering has also been useful in source code authorship attribution for detecting web spam [55] and exposing stylistic similarities in web forum posts [56].

### 2.1.2 Genre detection

Several studies focus on categorize texts based on their genre for various languages. Stamatatos et al. [57] utilized the same stylometric features in the genre detection like author attribution and verification tasks for modern Greek literature. Amasyalı and Diri [58] applied the word n-gram model with different machine learning classifiers for Turkish documents and achieved 93% accuracy. Sentiment and emotional sentence annotation are also utilized in predicting the genre of fictional text in several studies [59, 60]. Recently Kar et al. [61] proposed a hierarchical representation of narratives that improves over the traditional feature-based machine learning methods as well as sequential representation approaches for a multi-label dataset of narratives representing the story of movies. Worsham and Kalita [32] presented a study on how current deep learning models compared to traditional methods for this genre detection tasks in modern literature and discovered that an ensemble of chapters can significantly improve results.

### 2.1.3 Writing style change for authors

Authorial style of a writer who had a long career often changes significantly due to the change of language nature, experimenting new style, or due to the change in genre. However, there is a limited number of studies in this specific task of stylometry. Can and Patton [9] first investigated the changes in writing style over time of two Turkish authors using the average word length and most frequent words. Both the results from t-test and logistic regression showed that average word lengths of newer works were significantly larger than older works for both authors and therefore suggesting a correlation between word length and document age.

Klaussner and Vogel [10] developed a method of analysis that apply regression on linguistic variables in predicting a temporal variable changing over time for two English authors. Recently Gomex et al. [11] presented an approach to detect the change of writing style of seven authors for

three distinct phases. They have utilized various stylometric features (Phraseology, Punctuation, Lexical usage) to represent the novels in a vector space model and employed supervised learning algorithms to determine the writing stage of any particular author. The obtained result indicated that the stylometric feature might be a good solution for writing style change detection for some authors.

#### 2.1.4 Stylometry features

Researchers have identified stylometry features into five distinct categories: lexical, syntactic, semantic, structural, and domain (or content)-specific [51].

The most simple feature representation is lexical features. Lexical features are often word-based but sometimes can be character-based also. The bag-of-words (BoW) approach generally refers to lexical-level features as it represents a document as a bag (or collection) of words, discarding context, grammar, and word order. Most researchers rely on bag-of-words representations, given that lexical-level features typically yield state-of-the-art performance [31]. Researchers have also developed various vocabulary richness measures to quantify the diversity of sentences. These features can be both language-independent or dependent. Moreover, function words are also prevalent as stylometry features and have continued to be utilized in present researches [6].

Syntactic features capture patterns from the form of sentences. Syntactic data, while language-dependent, is highly reliable, assuming that precise and robust tokenizers, parsers, and part-of-speech taggers are usable. However, noise is added when these instruments are obsolete. Tags, phrasing, and rewriting rules are all syntactic characteristics that reflect the particular way in which sentences are structured by an author [62]. Semantic features capture meaning behind words, phrases, and sentences, such as through synonyms analysis and semantic dependencies. Clark and Hanon [63], for example, suggests a particularly interesting approach that considers synonym-based features as style indicators for the author attribution. The authors argue that their method takes into account the context of terms, where the choices of an author are valuable in word selection. However, these features are heavily language dependent and it requires the rigorous techniques to extract them from text [31].

Structural features define the organization of a document, such as how an author prefers to use indentations or signatures [64]. In online contexts, structural features are very useful, especially when structure is an important component of the document. Domain-specific features are often referred to as content-specific characteristics, because they depend on the document's content [64]. In this sense, the thematic and contextual clues provided in the document are included in the content.

The efficacy of these features depends on the characteristics of the dataset. Also, some features are not suitable for the literary domain and hard to extract for a low resource language like Bengali. Therefore, we aim to develop a new feature set to improve various literary tasks in the



Bengali language.

## 2.2 Character Interaction Graph and Application

In the last few decades, information visualization techniques have provided a new dimension in utilizing textual data. A character network is a graph extracted from a narrative, in which vertices represent characters and edges correspond to interactions between them. A number of narrative-related problems can be addressed automatically through the analysis of character networks. First, we discuss the overall works on the character interaction networks. Later, we provide further analysis on its use in social interaction extraction from stories.

### 2.2.1 Character Interaction Network

Character interaction representations are widely applied in the digital humanities to illustrate the relationships between characters in literary texts. Most of the works on character interaction focus on visualizing depict the social network between characters as a node-link static or dynamic graph [46]. Several researches also take advantage of this to solve higher-level problems such as, role detection [65], genre classification [66, 67], storyline detection [68], story segmentation [69], and others. They even reach the mainstream audience, specifically for their significance as a visualization tool for popular culture works.

We observe several variations and observations of character networks. Elson et al. [70] derived a graph network from dialogue interactions in nineteenth-century British novels and serials. In the network, vertices represent characters, and edges signify the amount of bilateral conversation between those pairs of characters, with corresponding edge weights proportionate to the frequency and length of their exchanges. Later Elsner proposed a kernel that evaluates the similarity between two novels in terms of the characters and their relationships, constructing functional analogies between them [71]. Ardanuy and Sporleder focus on building static and dynamic social networks of characters to represent the narrative structure of novels from different genres and authors [67]. For each novel, they have computed a vector of literary-motivated features extracted from their network representation and performed EM clustering in terms of genres and authorship.

Apart from literature, character interaction graph has also gained popularity in other media such as film, drama, tv-series, pop culture. Researchers have applied this approach to analyze movies, exploring new insights about specific movies and film industry in general. Character network from movies has been extracted either from screenplay processing [69, 72], subtitles [47], or script processing [73, 74]. The major distinction between different character interaction work lies in character identification, interaction detection, graph creation, and scope of application [46].

## 2.2.2 Social Interaction Analysis from Character Network Graph

Several studies aim to determine whether social interactions extracted from a fictional narrative display topological properties similar to real-world social networks. However, most of them concentrate on films and popular pop culture since character attributions and other necessary information are readily available online (IMDb, TDb, etc.) and can be extracted automatically. Stiller et al. [75] analyzed a corpus of ten plays by Shakespeare in an attempt to decide whether the success of this playwright depends on his ability to imitate certain fundamental properties of real social networks in his fiction. Alberich et al. [76] and Gleiser [77] evaluate how realistic cooperation between Marvel characters is by comparing their co-occurrence network with real-world collaboration networks. Using standard topological descriptors (degree distribution, average distance, transitivity, weight, and others), these studies define their character networks and compare their counterparts obtained for interaction in the real world.

Character interaction graph has also been utilized to answer various social science questions of contemporary times. Lauzen and Dozier [48] discuss the portrayals of different age groups and gender roles in top-grossing Hollywood films. They observe that both older men and women are dramatically underrepresented compared to their representation in real-life. In another work, Jarrott, and Mccann [25] analyzed 20 contemporary adolescent novels with intergenerational relationships using contact theory to assess whether the relationships between different age groups demonstrate attitudinal change. Recently Kagan et al. [47] investigated gender bias in on-screen female characters over the past century using a huge corpus of movie social networks. They discovered a trend of improvement in all aspects of women's roles in movies, including a constant rise in the centrality of female characters. Most of these researches include creating a character interaction graph from the narrative, incorporate different attributes to characters, and finally explore the trend using data analytic techniques.

## 2.3 Bengali Literature Analysis

Bengali, being one of the most spoken languages of the world, has an enriched literature history. However, there is a limited amount of studies in analyzing these literary texts using natural language processing (NLP) or data analytic techniques. We first discuss the existing stylometry tasks in Bengali literature. We also focus on the critical analysis Bengali literature and identify the scope of character network in literature analysis.

### 2.3.1 Stylometry Tasks in Bengali Literature

There is a limited number of studies in Bengali literature related NLP tasks and most of them focus on the author attribution. Das and Mitra [34] first investigated the author attribution

task in Bengali literary works on three authors and observed that simple unigram and bi-gram features along with vocabulary richness are strong enough to discriminate amongst these authors. Most of the following works pursued the same word/character n-gram [35, 78] and stylometry features [37, 79] to represent the text and applied machine learning algorithms Support vector machine (SVM), Neural Network to solve the classification task on the writings of newspaper columns or blog articles. Recently Chowdhury et al. [36] have investigated the effects and performance of word embedding models with deep neural networks for authorship attribution in Bengali. The performed experiments on a dataset of 2400 online blog articles from six authors revealed that skip-gram word embeddings by fastText tend to perform better than embeddings by Word2Vec or Glove.

### 2.3.2 Critical Analysis in Bengali Literature

There has been many critical analysis of the writings of prominent authors from a literature perspective. Chaudhuri [40] address the social changes reflected in Bengali literature fiction in the late nineteenth and early twentieth century where the portrayal of characters against a social standard is found more prominent. The role of women, gender, historical events have also been discussed in Bengali literature. Chatterjee [39] endeavors to a detailed study of the women characters in the novels of three nineteenth-century Bengali writers against the backdrop of the Indian Renaissance. The role of different religions and the perspective of authors are reviewed in several studies also [41, 42]. Other notable researches discuss national iconography of historical events [80], post-colonialism effects [81] in Bengali literature. Although we can have a clear picture of the social set-up in which those events occur from the characters depicted live, move and have their being [40], none of the prior studies attempt to explore this using a quantitative perspective.

### 2.3.3 Character Network Analysis in Bengali Fiction

The only prior study that focuses on character interaction in Bengali literature is Muhuri et al. [82]. They have extracted character networks from two plays of Rabindranath Tagore and proposed a novel idea to analyze the characteristics of protagonist and antagonist from the influential nodes based on the complex graph. However, their study does not explain the role of contemporary social set up or gender/age group effect in the character interaction. Also, analysis of only two works of an author does not provide us a detailed overview of the author's perspective on social issues.

From our literature overview, it is evident that the existing studies in Bengali literature do not provide an analysis on how writing styles evolve for an author or which features contribute primarily to authorship distinction. Therefore, we utilize *Word2vec graph* model features

along with word unigram and stylometry feature sets to explain these queries as well as author identification, genre detection, and stylochronometry tasks. Also, our study is the first quantitative assessment in Bengali literature analysis that exploits the character interaction graph to alleviate the limitations of existing manual critical analysis works.

## Chapter 3

# *Word2vec graph* Model for Stylometry Tasks and Literary Analysis

In this chapter, we propose a novel word2vec graph based modeling of a story that can rightly capture both the context and the structure of the story. By using these word2vec graph based features, we develop a classification technique to perform author attribution, genre detection, and stylochronometry tasks. First, we explore our proposed *Word2vec graph* model. Then we describe the experimental evaluation involving the dataset creation and baseline methods discussion. Finally, we present our results and overall analysis.

### 3.1 The *Word2vec Graph* Model

First, we describe the process of constructing a *Word2vec graph* from a document and then discuss the feature extraction from the *Word2vec graph*. After that, we augment *Word2vec graph* feature set with corresponding words features, and devise a clustering technique based on these features and words. This clustering technique will be subsequently used in author identification and genre identification task. An overview of our proposed system is presented in Figure 3.1.

#### 3.1.1 The *Word2vec graph* creation

Word2vec is a two-layer neural net, the most common word embedding technique that represents a fixed vector size of every word in the corpus [83]. We utilize this word2vec embeddings to represent the similarities between words in a document as a graph. Each node in the *Word2vec graph* denotes a word and edge between nodes represent their relation in the document. We consider edge weight as the cosine similarity between vector representations of corresponding words. The concept of *Word2vec graph* is motivated from the word similarity graph [84, 85]. For the whole corpus, a single word similarity graph is created, and the similarity between

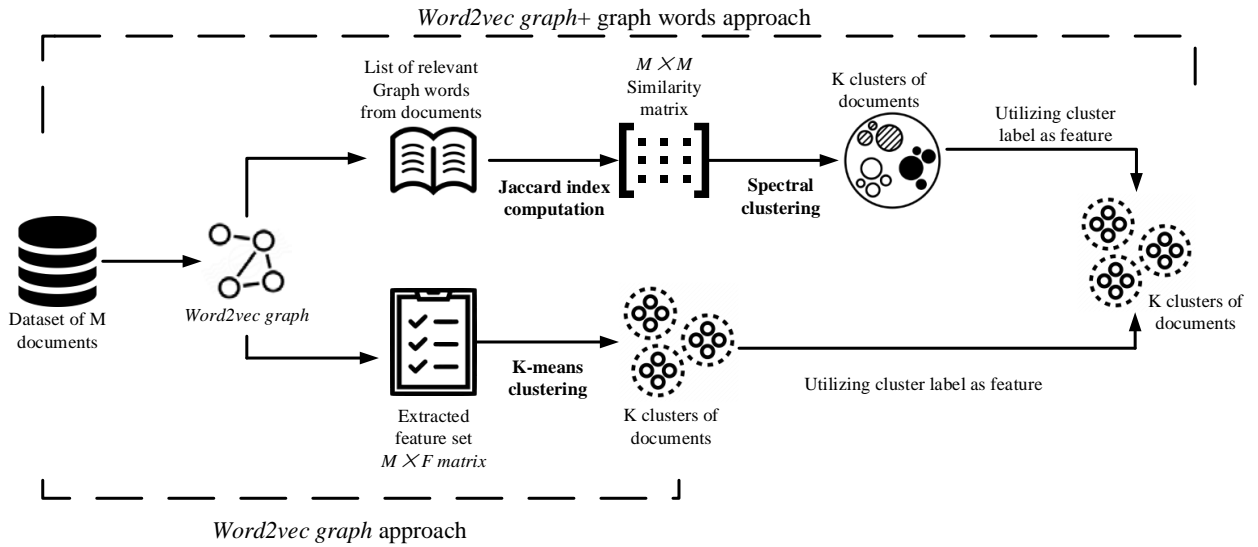


Figure 3.1: Overall architecture of *Word2vec graph* and *Word2vec graph+Graph words approach*

words is measured by their co-occurrence in documents. However, in our problem, we represent each document as a *Word2vec graph*, and the similarity between words is measured from the word2vec vector representations. Word similarity graph is commonly used for community detection problems to find related groups of words/synsets, where *Word2vec graph* is used to classify a document.

To generate a *Word2vec graph* from a sample document, we first train a word2vec model from the tokenized format of the document. This model transform each word  $w$  in the document to with a dense vector  $\bar{w}$  with  $l$  dimension (we consider  $l = 100$ ). We take top  $N$  words (*core words*) in the document with most frequency. For each *core* word  $i$ , we consider  $K$  most similar words in the documents. We calculate the cosine similarity between each of these similar words  $j$  and corresponding core word  $i$ , which represent the edge weight  $\omega_{ij}$ . We also consider each node weight  $w_i$ , as the relative frequency of that word in the document. We categorize nodes in the *Word2vec graph* into three types according to their relative index based on frequency in document and connectivity with other nodes. Similarly, edges are also classified by their connectivity to different nodes.

- **Core nodes, edges:** Corresponding nodes of top- $N$  words in the document. Any edge among core nodes is represented as the *core edge*.
- **Multiple nodes, edges:** The nodes which are connected to two or more core nodes in the graph. The edge between a multiple node and core node is represented as *multiple edge*.
- **Boundary nodes, edges:** The nodes which are connected to only one core node in the graph. The edge between a multiple node and core node is represented as *boundary edge*.

While creating *Word2vec graph*, we also address the effect of function words/stopwords since they play an effective role in various NLP task [86]. However, the relative frequency of function words/stopwords is higher than the context words in many cases, which might decline the efficacy of *Word2vec graph*. Therefore, we consider both versions of *Word2vec graph*, one *with-stopwords*, and other *without-stopwords* for each document. We utilize a Bengali stopword list from a publicly available source<sup>1</sup> and improve it by inserting missing words during experimentation. However, during experimentation, it becomes evident that it does not consider many relevant stopwords in our corpus, especially from *Sadhu-Bhasha* dialect. Therefore, we manually include those words from our text and improve the stopword list. However, some stopwords such as 'করা' can be used with verbs and create different meanings for these verbs. It can be considered as a stylistic signature of authors. Therefore, we remove them from our stopwords list.

Apart from stopwords removal in the *without-stopwords* version, we do not perform any other pre-processing tasks, such as lemmatization/stemming, to keep the intact form of words. We consider  $N = 20$  and  $N = 15$  for *with-stopwords* and *without-stopwords* version respectively and  $K = 10$  in both cases to keep identical structures for different documents.

### 3.1.2 Features extraction from *Word2vec graph*:

Different graph embedding methods have been popular recently to represent the graph structure, such as DeepWalk [87], Graph Convolution Network [88], Graph2vec [89]. However, we utilize some pre-defined features to represent *Word2vec graph* since the distinction and connectivity between core, multiple, and boundary nodes/edges provide clear insight into the graph in this scenario. Besides, these features can capture the representations of corresponding text and relativeness between words. Therefore, we utilize a set of key features to represent *Word2vec graph* structure for simplicity. Apart from count, weight, degree of various nodes/edges, we also consider the relative index of nodes (ordered index of words based on frequency). There is a total of 95 and 85 features respectively, representing the graph structure for *with-stopwords* and *without-stopwords* version. Table 3.1 provides a list of features that we extract from *Word2vec graph* of stories in both versions.

### 3.1.3 Clustering with *Word2vec graph* features:

We prefer clustering rather than conventional classification methods, as many writers frequently write on the same subject/ genre, and some of their literary works can reflect close similarities in terms of styles or contexts. Clustering derives a natural separation of the feature space that may or may not correlate with the class labels, and can be used to identify stylistic similarities [31].

<sup>1</sup><https://github.com/stopwords-iso/stopwords-bn>

Table 3.1: Extracted features from *Word2vec graph*

Feature	#
No. of core/multiple/boundary nodes/edges	6
Min/max/avg/sum of core/multiple/boundary nodes/edges weights	18
Count/min/max/avg of core nodes degree (considering core/multiple/boundary edge only)	10
Degree count of core nodes (considering core/multiple edge only)	30/40
Min/max/avg/stdv index of multiple/boundary nodes	8
No. of multiple/boundary nodes index under threshold	2
Core nodes having degree 0-5 (considering multiple edges only)	6
No. of core nodes having degree equal to min/max/greater than avg/smaller than avg	4
<b>Total features</b>	<b>85/95</b>

Therefore, even if the predicted cluster is inaccurate for any author or genre, these clusters may depict the similar writings of various authors.

Given a dataset of  $M$  documents (stories/articles), we generate *Word2vec graph* for each document. We extract relevant features from the graphs to convert the dataset as a  $M \times F$  matrix, where  $F$  is the number of features. Then we perform k-means clustering on it to categorize documents into  $k$  groups. For any particular experiment, we perform clustering with a fixed number of clusters (No. of clusters  $k$  equals to distinct author/genre present in the dataset used for that experiment). Since this feature list does not contain any actual words relevant information (indexes of words vary for each document), it might miss some contextual information of the document. Therefore, we utilize combinations of the *core*, *multiple*, and *boundary* words information. These words represent the most frequent words in the document, along with their associated words. Thus, they are capable of obtaining context from the overall narrative of the document.

Given two document  $a, b$  and their corresponding graph words set  $W_a, W_b$ , we compute their Jaccard similarity coefficient [90] as,  $Sim(W_a, W_b) = (W_a \cap W_b)/(W_a \cup W_b)$ . We generate a similarity matrix  $M \times M$  from each pair of documents in the dataset and perform Spectral clustering [91] with  $k$  clusters. Spectral clustering is particularly useful to find clusters when a similarity matrix is provided and uses information from the eigenvalues (spectrum) of special matrices built from the similarity matrix. Finally, we combine this clustering result with previous *Word2vec graph* result as features and perform another k-means clustering to generate the final cluster information. We denote this approach as *Word2vec graph+graph words*.

## 3.2 Experimental Evaluation

We provide a brief description on corpus generation procedure and discuss the baseline methods in this section.



### 3.2.1 Corpus Creation

There are no publicly available datasets for Bengali literature analysis. Although there is some newspaper corpus available in the Bengali language, they contain news from various topics/tags. Therefore, it is not suitable to study author style or attribution task.

Since Bengali exhibits two major forms of dialects, two styles of writing have emerged, and that involves somewhat different vocabularies and syntax [23]. *Sadhu-Bhasha* or chaste language is an old written formal style of Bengali language, with longer verb inflections and more of a Pali and Sanskrit-derived vocabulary. *Cholito-Bhasha*, known by linguists as standard colloquial Bengali, is a written Bengali style that is comparatively easy and informal. It exhibits a preponderance of colloquial idiom and shortened verb forms. Modern Bengali literature advanced in *Sadhu-Bhasha*, and it was widely used for Bengali prose till the third decade of the 19th century. However, *Sadhu-Bhasha* in modern writing is extremely rare, and it is restricted to some official signs and documents in Bangladesh as well as for achieving particular literary effects. Colloquial language is now generally used in both formal and informal writings. Moreover, the writing form, syntax or choice of verbs/phrases differ greatly based on the origin, such as Bangladesh or West-Bengal. Therefore, we try to incorporate these varieties during corpus creation.

Table 3.2: Different authors with their characteristics and book counts on various genres in Bengali literature corpus.

Author Name	Origin	Career	Dialect	Collected Book Counts
Bankim Chandra (BC)	Bengali Province	1866-1879	<i>Sadhu</i>	Novel: 14
Rabindranath Tagore (RT)	Bengali Province	1883-1940	<i>Sadhu, Cholito</i>	Novel: 12, Short story: 104
Sarat Chandra Chattopadhyay (SCC)	Bengali Province	1914-1931	<i>Sadhu</i>	Novel: 24, Short story: 29
Humayun Ahmed (HA)	Bangladesh	1970-2011	<i>Cholito</i>	Novel: 103, Long/Historical novel: 7, Himu series: 23, Misir Ali series: 20, Short story: 110, Science fiction: 10
Sunil Gangopadhyay (SG)	West Bengal	1965-2012	<i>Cholito</i>	Novel: 8, Long/Historical novel: 6, Kakabau/Thriller series: 36
Shirshendu Mukhopadhyay (SM)	West Bengal	1967-	<i>Cholito</i>	Novel: 4, Long/Historical novel: 3, Thriller: 7, Teenage novel: 39
Samaresh Majumdar (SMM)	West Bengal	1976-	<i>Cholito</i>	Novel: 8, Short story: 72
Muhammed Zafar Iqbal (ZI)	Bangladesh	1982-	<i>Cholito</i>	Teenage novel: 11, Science fiction: 33

### Bengali Literature Corpus

We create a corpus containing the writings of eight distinguished authors in Bengali literature. Our corpus contains relevant information for each story, such as story title, book type, chapter title, chapter text. The authors are from distinct periods, and we try to compile as much as writings possible from different genres. The authors are from Bangladesh, previous Bengali province in the Indian subcontinent, or the current West-Bengal province in India. A summary of the literature corpus is provided in Table 3.2.

### Newspaper Editorial Corpus

We create a newspaper corpus from five prominent newspapers <sup>2</sup> in the Bengali language, such as Prothom Alo, Ittefaq, Jugantor, Inqilab, and Anandabazar. The major focus of this study is to find the efficacy of different feature sets in exploring author similarity. Therefore, we only consider the editorial writings since they are supposed to be written by the editors/sub-editors of the newspaper and should contain the writing signatures of them. We crawl from these newspaper pages with the editorial tag and retrieve relevant information, such as publishing date, author name (if available), sub-tags. Since editorial pages also include letters to the editors and various articles by prominent scholars of the country, we remove them to make the dataset less diversified. Finally, to reduce the topic and context dissimilarity among these articles, we crawl between a fixed period (May 2019 - November 2019) for all newspapers. Table 3.3 provides a summary of the newspaper editorial corpus.

Table 3.3: Article counts and characteristics of different newspaper editorials

Newspaper Name	Origin	Dialect	Article Count
Prothom Alo	Bangladesh	<i>Cholito</i>	254
Ittefaq	Bangladesh	<i>Sadhu</i>	254
Jugantor	Bangladesh	<i>Cholito</i>	222
Inqilab	Bangladesh	<i>Cholito</i>	202
Anandabazar	West Bengal	<i>Cholito</i>	257

### Project Gutenberg Corpus

To show the effectiveness of *Word2vec graph*, we also perform various stylometry tasks on a subset of the Project Gutenberg corpus. Project Gutenberg is an extensive web catalog containing over fifty thousand e-books. We utilize a small subset of the Project Gutenberg corpus from [92]. All books have been manually cleaned to remove metadata, license information, and transcribers' notes, as much as possible. Along with providing the text for all of these books, Project Gutenberg

<sup>2</sup><https://www.prothomalo.com/>, <https://www.ittefaq.com.bd/>, <https://www.jugantor.com/>, <https://www.dailyinqilab.com/>, <https://www.anandabazar.com/>

also reports a detailed index for each book which contains the title, author, publication date, and Library of Congress Subject Headers (LCSHs). Our version of the dataset consists of six categories of books from 54 authors: Science fiction, Adventure stories, Historical fiction, Love stories, Detective and mystery stories, and Western stories. Overview of our version of the Gutenberg dataset can be found in Table 3.4.

Table 3.4: Overview of our Gutenberg corpus

Genre	Count	Average # sentence	Mean Chapter Count
Detective and mystery stories	72	8695	18.4
Love stories	64	17691	25.1
Historical fiction	60	15496	27.9
Adventure stories	71	9696	24.5
Western stories	38	9955	22.9
Science fiction	46	6803	9.5

### 3.2.2 Baseline Methods

Content-based features tend to be suitable for datasets with high topical diversity where datasets with less topical variance, benefit more from style-based features [33]. Therefore, we consider both Bag of words (unigram) feature set having Term Frequency Inverse Document Frequency (TF-IDF) score to capture the content and stylometry feature set (lexical, syntactic, sentiment, etc.) to find the stylistic signature of authors and utilize them as baseline methods. Both of them are also effective in text classification tasks regardless of the domains [43].

#### Bag of Word Feature Set with TF-IDF Score

Bag of words is a primitive but still highly successful approach in document classification. We utilize the TF-IDF score of top  $N$  words in the vocabulary to represent a document (story/article). First, we generate a vocabulary of unique words from all documents in each corpus (Literature/Newspaper). Since the vocabulary size is large (95K for literature corpus, 260K for Gutenberg corpus and 53K for newspaper corpus) and many words have a very small frequency, we consider unigrams only to avoid overfitting and reduce feature dimension. Given  $|D|$  documents in corpus, we compute TF-IDF score for each word  $w$  in a document  $d$  as follows. Similar to *Word2vec graph*, we have considered both *with-stopwords* and *without-stopwords* version for the TF-IDF feature set.

$$tf - idf(w, d) = tf(w, d) \times \log(|D|/(df + 1))$$

Here, term frequency is  $tf(w, d) = \text{count of } w \text{ in } d / \text{number of words in } d$  and document frequency is  $df(w) = \text{occurrence of } w \text{ in documents}$ . Although we compute TF-IDF score for

all words in the vocabulary, we consider top  $F$  words in the corpus while creating the feature matrix  $M \times F$ . Then we perform k-means clustering algorithm like earlier approach. We take  $F = 30K$  for Bengali literature & Project Gutenberg corpus and  $F = 15K$  for newspaper corpus.

### Stylometry Feature Set

Stylometry feature set is extremely popular and widely utilized in various document classification sub-tasks. We mostly incorporate lexical, character-based features from each story/article since they are applicable to any language/corpus and suitable for an under-resourced language like Bengali. We also include some syntax features and parts of speech (pos-tag) information of words. Semantic analysis can be difficult even with language processors, particularly on unrestricted texts [31] like literature, and NLP in Bengali is still not sufficient to extract these features. Structural features are more effective in online text and not suitable for our corpus. However, we exploit the sentiment and emotion property of words to make the feature set more effective.

Bengali alphabet has 11 vowels, 39 consonants, and not distinct cases. Its vowel graphemes are mainly realized not as independent letters, but as diacritics, modifying the vowel inherent in the base letter where they are added. Sample Unicode based parser is capable of parsing or tokenizing Bengali characters/words. English words and digits are also present in the corpus, expressing any technical term, or quoting any statements. For parts of speech tagging of Bengali words, we have utilized an open-source corpora<sup>3</sup> that contain the corresponding pos-tag of nearly 100K Bengali words (many forms of the root words are present and only major pos-tag information, such as Noun, Pronoun, Adjective, Adverb, and Verb are available).

Existing works in sentiment and emotion prediction in Bengali language only focus on specific domains, such as Twitter [93], YouTube comments [94], blog posts [95], etc. Therefore, these approaches are not suitable to detect sentiment/emotion from sentences/words in our corpus. Therefore, we only consider the sentiment/emotion tag of words in the document using Bengali SentiWordNet [96] and WordNet Affect [97]. However, there is a limited number of words in both of them, and they do not cover a significant portion of our corpus. So, we further improve them by correcting several entries manually and incorporating more words from English language resources, such as SenticNet4 [98], AFINN [99], Multilingual WordNet Affect [61]. There are 9321 words in our version of SentiWordNet, and each has a score between -1 to +1. Our EmotionNet contains 1017 words, and each word has one or more emotion tags according to Ekman's [100] six emotions class (Joy, Anger, Disgust, Fear, Surprise, Sad).

There is a total of 223 features in the stylometry feature set. We employ a nearly similar feature set for Project Gutenberg corpus, replacing the frequency/count of Bengali alphabets/words with English. We provide some example features for a document in Table 3.5.

<sup>3</sup><https://github.com/sunku02/BanglaPosTagger>

Table 3.5: Extracted stylometry features

Type	Feature	#
Lexical (Character)	Relative frequency of Bengali characters/graphem/digits	~70
	Relative frequency of words starting with specific characters/graphem	~60
	Total percent of Bengali/English digit	2
Lexical (Word)	No. of total words/unique words/sentences	3
	Avg word/sentence length	2
	Relative frequency of words with length (1-20)	20
	Relative frequency of sentences with length (1-20)	20
Syntactic	Relative frequency of different pos-tags	6
	Avg no. of specific punctuation per sentence	~10
	Stopwords avg frequency per sentence/total document	2
Sentiment/ Emotion	Relative frequency of pos/neg/neutral words	3
	Percentage of sentence having pos/neg sentiment score	2
	Relative frequency of various emotion tags	6

### 3.3 Results and Discussion

In this section, we evaluate the performance of *Word2vec graph* feature set with baseline approaches in both Literature and Newspaper corpus for various tasks. We also identify the most important attributes in these feature sets that contribute primarily to these tasks. Finally, we discuss whether these stylometry signatures evolve significantly for authors who have a prolonged career.

Although we are performing unsupervised clustering, the original label information (author name/genre) is available for the instances. We obtain an optimum allocation of the stories from the predicted cluster label and the original label by utilizing the Hungarian Algorithm [101]. Since there is a class imbalance in the datasets, we adopt weighted F1 score as the performance metric. The F1 scores are calculated for each label and then their average is weighted by support - which is the number of true instances for each label.

For all feature sets (except stylometry), we conduct experiments with both *with-stopwords* and *without-stopwords* versions, and report the better one. We denote (s) to represent that better results were found with *with-stopwords* version for that feature set. We indicate the inclusion of *core*, *multiple*, and *boundary* words by using (c), (m), (b) respectively for *Word2vec graph+graph words* feature set. For example, (c, m) specifies that the best result for *Word2vec graph+graph words* feature set was found by adding *core* and *multiple* words and using the *without-stopwords* version.

#### 3.3.1 Result on Bengali Literature Corpus

we perform our adopted clustering methods by using various feature sets for author identification, genre detection, and determining writing phases for authors on Literature corpus. We conduct

several experiments with various combinations of authors/genres and present the summary of results.

### Author Attribution

Table 3.6 shows the overall results for the author identification problem in different scenarios. Each criterion represents a portion of the dataset that was utilized for that particular experiment. For example, "Novel, short story (HM, RT, SC, SMM)" specifies that we consider only the novels and short stories of four authors and perform k-means clustering using four clusters.

Table 3.6: Weighted F1 score for various feature set in author identification task.

Criteria	Sample #	w2v	w2v+words	TF-IDF	Stylo
Science fiction (HM, ZI)	43 (k=2)	<b>1</b>	<b>1 (c)</b>	<b>1</b>	.887
Teenage novel (SM, ZI)	50 (k=2)	.916	.978 (c) (s)	<b>1</b>	.879
Long historical novel (HM, SG, SM)	16 (k=3)	.838	<b>1 (c) (s)</b>	<b>1</b>	.789
Series novel (HM, SG)	79 (k=2)	.738	<b>1</b>	<b>1</b>	.713
Thriller (SG, SM)	43 (k=2)	.84 (s)	<b>1</b>	.979 (s)	.761
Short story (HM, RT, SC, SMM)	308 (k=4)	.457 (s)	.574 (c) (s)	<b>.889 (s)</b>	.343
Novel (RT, BC, SC)	50 (k=3)	.512	.915 (c, m)	<b>1</b>	.416
Novel (SG, SM, SMM)	20 (k=3)	.691	.801 (c, b)	<b>1</b>	.463
Novel, short story (HM, RT, SC, SMM)	462 (k=4)	.482	.567 (c) (s)	<b>.606</b>	.397
Novel, historical novel (HM, SG, SM)	131 (k=3)	.780	.844 (c)	<b>.932</b>	.567

TF-IDF feature set achieves the highest F1 score in nearly all cases. *Word2vec graph+graph words* performs almost equal in most scenarios. *Word2vec graph* and Stylometry feature set specifically struggle where short stories are involved. Because of the limited size of data, *Word2vec graph* can not properly capture the representation of the story. Both TF-IDF and *Word2vec graph* feature sets perform better in *without-stopwords* version. However, stopwords play an important role in author identification especially in specific cases when short stories are present. Because removing these words further shrinks the size of texts and reduces the performance of feature sets. Adding graph words can enhance the performance of *Word2vec graph* in author identification. However, *boundary* words are not effective since it is evident that *Word2vec graph+graph words* performs better with the inclusion of *core* words only or a combination of *core* and *multiple* words.

Figure 3.2, 3.3 shows the clustering visualization on the total corpus (except short stories since they all form a separate cluster combining the writings of respective authors) with a fixed number of clusters using TF-IDF feature set in *with-stopwords* and *without-stopwords* versions respectively. For *with-stopwords* version, if we consider two clusters, writings of Bankim, Sarat, Rabindranath (partially) form one cluster (stories in *Sadhu-Bhasha*). The second cluster constitutes of rest of the writings of Rabindranath and other authors (stories in *Cholito-Bhasha*). If we increase the number of clusters (for example, four), writings of Humayun Ahmed and

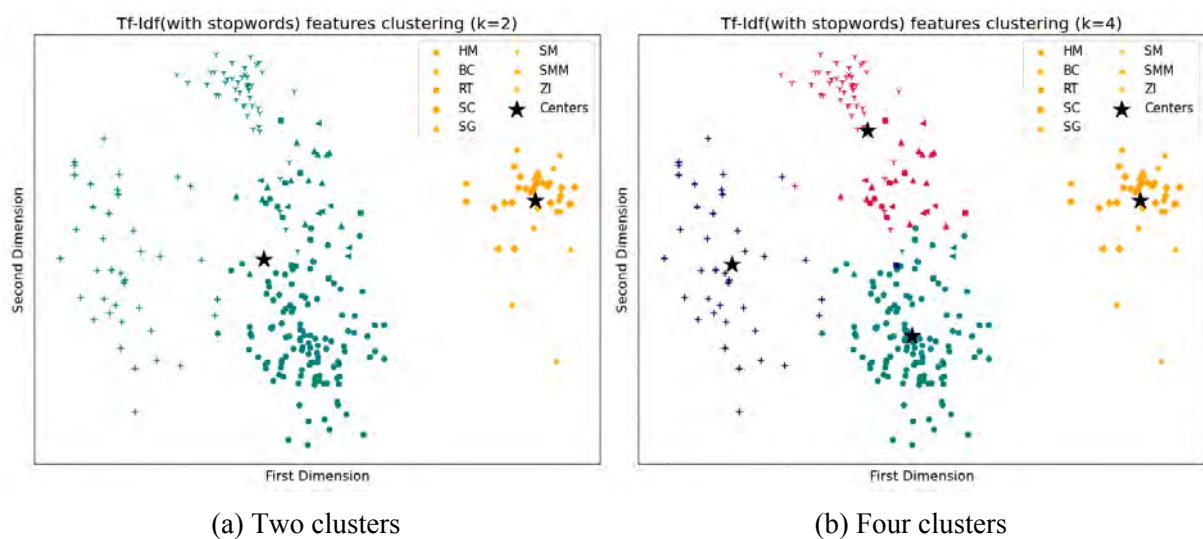


Figure 3.2: k-means clustering visualization for all authors using TF-IDF feature set *with-stopwords* version. Every point denotes a sample story. Each color represents a cluster and each marker represents the writing of an author.

Muhammed Zafar Iqbal form their separate clusters (Bangladeshi authors), and writings of Sunil, Shirshendu, Samaresh, and partially Rabindranath (West Bengal and previous Bengali authors who wrote in *Cholito-Bhasha*) constitute a single cluster. Similarly, for *without-stopwords* version, we get a cluster of Humayun stories and second cluster formed by the stories of other authors, if we consider  $k = 2$ . We observe similar results (like *with-stopwords* version) by increasing the cluster number to four.

These results demonstrate that word usages of previous authors in Bengali literature follow a specific pattern that is not observable in current days authors. Humayun Ahmed portrays a specific writing signature by his unique word patterns. West Bengal authors often show common word usages in their writings. We find almost similar results by applying *Word2vec graph+graph words* feature set. Therefore, even if having a significantly lower number of features, *Word2vec graph* with graph words has the capability to represent author patterns from stories.

### Genre Detection

We perform the genre detection task on the writings of single and multiple authors using various combinations. We present some significant cases in Table 3.7.

*Word2vec graph+graph words* performs better in most of the scenarios. *Word2vec graph* achieve nearly equal F1 score in these cases. TF-IDF feature set is successful, especially in distinguishing series writings (*Himu* and *Misir Ali* series) of Humayun Ahmed from normal novels. Each series contains the same set of characters in different books, and so word usage follows specific patterns. Stylometry feature set performs best when short stories are involved. All short stories have some specific stylometry features, such as total sentences/ words that make it easily separable from

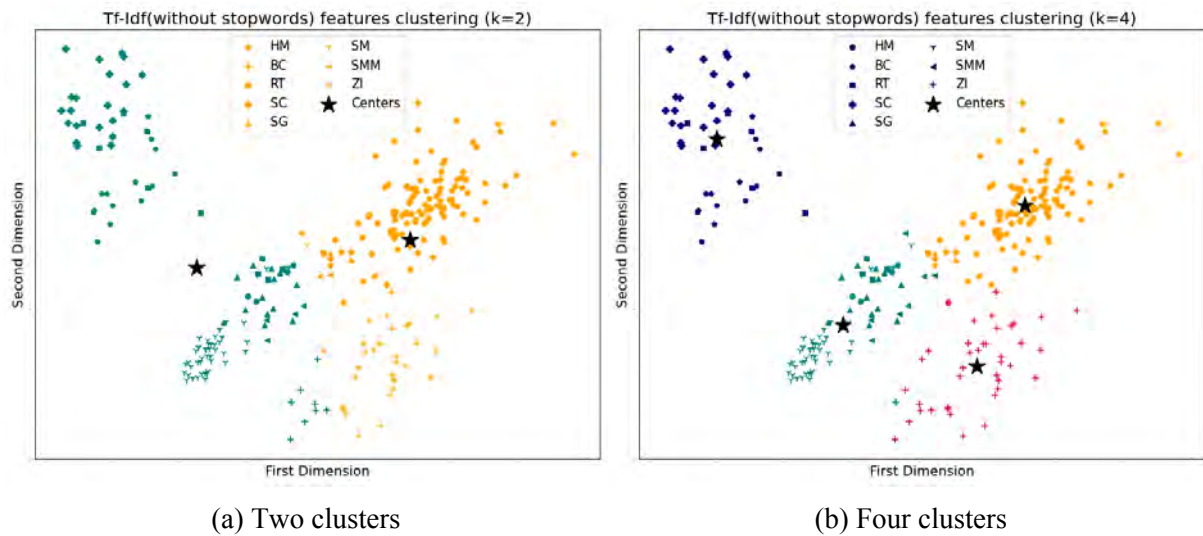


Figure 3.3: k-means clustering visualization for all authors using TF-IDF feature set *without-stopwords* version. Every point denotes a sample story. Each color represents a cluster and each marker represents the writing of an author.

Table 3.7: Weighted F1 score for various feature set in genre detection task

Criteria	Sample #	w2v	w2v+words	TF-IDF	Stylo
HM, RT, SC, SMM(novel, short story)	462 (k=2)	.794 (s)	.794 (c) (s)	.626	<b>.805</b>
HM, SG, SMM(novel, historical novel)	131 (k=2)	<b>.936</b>	<b>.936 (c)</b>	.676	.785
HM(novel, Himu, Misir Ali)	146 (k=3)	.431	.664 (c)	<b>.893</b>	.519
HM(novel, science fiction)	113 (k=2)	.751	<b>.782 (c) (s)</b>	.689	.698
SM(teenage novel, thriller)	46 (k=2)	.767	<b>.962 (c) (s)</b>	.825	.767
ZI(teenage novel, science fiction)	44 (k=2)	<b>.977</b>	<b>.977 (c, m)</b>	.583	.646

other genres. However, it is evident that *Word2vec graph* by adding graph words can successfully represent the genre/content information of the stories.

Stopwords/function words do not contain any topic/genre-related information. So, TF-IDF feature set works well in *without-stopwords* version. In contrast, the relation of these function words with content words may be beneficial because *Word2vec graph+graph words without-stopwords* performs better in identifying genres like novel, historical novel. Using *core* words as graph words are sufficient in most of the cases. Although, for some specific genres, such as science fiction, teenage novel, thriller, *Word2vec graph+graph words* work better by adding *multiple* words. These words often contain some genre/ content information.

### Clustering based on time

Although there are 700 stories in our corpus, we only consider novels since they can provide better insight regarding the stylometry developments of authors. Also, we are particularly interested in those authors whose literary career spans for a longer time and overlaps with others. We assemble writing years for five authors as depicted in Table 3.8. Since Humayun Ahmed and



Sunil Gangopadhyay had a contemporary career of writing, we separate their writings into two (writings before and after 1995) and three (writings before 1990, 1990-2000 and after 2000) groups for evaluation purposes. For other writers, we consider two groups. We try to ensure that these groups have a nearly equal number of stories. Table 3.8 shows the k-means clustering performance using various feature sets on different criteria.

Table 3.8: Performance of various approach in clustering wrt timing of writing

Criteria	Sample #	w2v	w2v+words	TF-IDF	Stylometry
HM(<1995, >1995)	74 (k=2)	.608 (s)	.621 (c) (s)	<b>.765 (s)</b>	.554
HM(<1990, 90-00, >2000)	74 (k=3)	.399 (s)	.434 (c) (s)	<b>.706</b>	.394
SG(<1995, >1995)	33 (k=2)	.666 (s)	.816 (c)	<b>1</b>	.606
SG(<1990, 90-00, >2000)	33 (k=3)	.560	.764 (c)	<b>.935</b>	.515
BC(<1875, >1875)	14 (k=2)	.642 (s)	<b>.857 (c) (s)</b>	<b>.857 (s)</b>	.615
RT(<1917, >1917)	11 (k=2)	<b>.909 (s)</b>	<b>.909 (s)</b>	<b>.909 (s)</b>	.605
SC(<1917, >1917)	21 (k=2)	.805 (s)	.809 (c, m) (s)	<b>.904 (s)</b>	.761

TF-IDF feature set performs most in all cases, and *Word2vec graph+graph words* performs almost equally and achieves a higher F1 score than *Word2vec graph* and Stylometry feature sets. Therefore, it is evident that word usage plays an important role in evolving the writings of an author over time. The highest performance is observed for Rabindranath Tagore and Sunil Gangopadhyay's writings. Tagore utilized *Sadhu-Bhasha* his earlier works and *Cholito-Bhasha* in later stories. Kakababu series of Sunil Gangopadhyay introduced new characters/storylines in each book, and recent books utilize contemporary words that are easily distinguishable from earlier books. Therefore, TF-IDF and *Word2vec graph+graph words* have a higher F1 score than other scenarios.

The first few novels of Bankim Chandra were experimental in Bengali literature and were highly influenced by Sanskrit language. Later stages novels of Sarat Chandra were also inspired by recent trends of that period (utilizing more character-oriented plots, description of places/society) [102]. Therefore, comparative performance is observed for the writings of both of these authors. It is also observed that *with-stopwords* version performs better for earlier authors like Tagore, Bakim, and Sarat.

The least performance of all feature sets is observed for Humayun Ahmed's writings. Humayun Ahmed had a versatile writing career and introduced a different level of stylometry in all his writings. His writings are easily noticeable from other authors as discussed in the earlier subsection. But novels of Humayun Ahmed contain almost similar word usage and writing patterns. Therefore, it is difficult to categorize them based on the timing, and the performance of all feature sets is significantly less than other authors.

### 3.3.2 Result on Subset of Project Gutenberg Corpus

Since author information and book categories are available for our version of the Gutenberg dataset, we perform both author attribution and genre detection.

#### Author Attribution

Table 3.9 shows the overall results for the author attribution problem in different scenarios. Each criterion represents a portion of the dataset that was utilized for that particular experiment. We perform author attribution in each genre and all books. However, we consider only the authors with books count greater than ten to remove outliers and potential noisy samples in clustering.

Table 3.9: Result on author attribution in Project Gutenberg Corpus

Criteria	Sample #	w2v	w2v+words	TF-IDF	Stylometry
Detective and mystery stories	52 (k=2)	.856	<b>1</b>	.9244	.878
Love stories	17 (k=2)	.646	<b>1</b>	<b>1</b>	.684
Historical fiction	19 (k=2)	.578	<b>1</b>	.947	.526
Adventure stories	39 (k=4)	.435	<b>.861</b>	.713	.461
Western stories	27 (k=2)	.782	.712	<b>.775 (s)</b>	.769
Science fiction	28 (k=3)	.720	<b>.854</b>	.776	.758
All genres	201 (k=6)	.504	.681	<b>.696</b>	.453

It is evident that our method *Word2vec graph + word* method achieves higher performance in most of the scenarios. Also, the *without-stopwords* version works better since, in English literature writings, the most frequent words of all authors are the stopwords and nearly similar. Therefore, they do not provide any significant information regarding author characteristics and reduce the performance. TF-IDF performs better especially in author attribution in western story genres. Authors of these books often use special, localized, regional words. Therefore, the limited number of words considered in *Word2vec graph+word* method can not capture them all. The *Word2vec graph* also achieves similar or better than the stylometry feature set in most of the cases.

#### Genre Detection

We perform genre detection on all authors. We also consider authors with more than ten books to observe the effect of outliers and noisy samples during clustering. It is interesting that the *Word2vec graph* structure performs best if we consider all authors. *Word2vec graph + word* performs best after removing the outliers and the F1 score also improves significantly. Both of them are better than the baseline TF-IDF and stylometry methods.

Table 3.10: Result on genre detection in Project Gutenberg Corpus

Criteria	Sample #	<i>w2v graph</i>	<i>w2v+graph words</i>	TF-IDF	Stylometry
All authors	334 (k=6)	<b>.365</b>	.337	.257	.341
All authors with book count $\geq 10$	192 (k=6)	.391	<b>.432</b>	.416	.396

### 3.3.3 Result Analysis on Newspaper Editorial Corpus

To show the effectiveness of feature sets in another domain, we perform k-means clustering in newspaper editorial articles to obtain clusters and evaluate whether they belong to the same newspaper. However, editorial articles contain discussion on distinct topics, and that can be crucial for the performance of feature sets. Therefore, we utilize the approach in [33] to calculate the topical similarity among newspapers. We use Latent Dirichlet Allocation (LDA) [103] to generate a topic distribution of the articles where each topic is characterized by a distribution over words. Let  $C_\alpha$  is the set of articles in newspaper  $\alpha$  and  $\phi_i$  is the topic distribution for the  $i$ -th document in  $C_\alpha$ , then we estimate the topic distribution for a newspaper editorial writings,  $\theta_\alpha$ , as follows:

$$\theta_\alpha = \frac{\sum_{i=1}^{|C_\alpha|} \phi_i}{|C_\alpha|}$$

Finally, we use Jensen-Shannon Divergence (JSD) [104] to calculate the difference between two newspaper's topic probability distributions. High JSD scores indicate a more topical dissimilarity between two newspapers. A matrix of JSD scores between newspapers is provided in Figure 3.4a for 20 topics. Similar results were found by changing topics number. The highest topical dissimilarity is found for the Anandabazar newspaper since it is a West-Bengal newspaper and certainly covers different topics. Ittefaq has also higher JSD value since it contains articles in *Sadhu-Bhasha*.

Table 3.11 shows the results for different feature sets. We perform clustering on different criteria, such as all newspaper articles with *Cholito-Bhasha* (An, In, Ju, Pr), all Bengali newspaper with *Cholito-Bhasha* (In, Ju, Pr), all Bengali newspaper (It, In, Ju, Pr), whole corpus, etc. In most of the cases, *Word2vec graph+graph words* or *Word2vec graph* feature set achieve highest performance. TF-IDF only performs better with the inclusion of *Sadhu-Bhasha* newspaper articles (Ittefaq). TF-IDF feature set has less F1 score when topical dissimilarity between newspaper articles is low. Adding *core* and *multiple* words with *Word2vec graph* features can increase the performance when there is a certain topical dissimilarity between articles. Otherwise, *Word2vec graph* performs best when clustering is performed on the total corpus. Therefore, *Word2vec graph* features can successfully portray the writing signature of editorial articles when the topic or word usage information is not effective. In all scenarios, *with-stopwords* version of feature sets

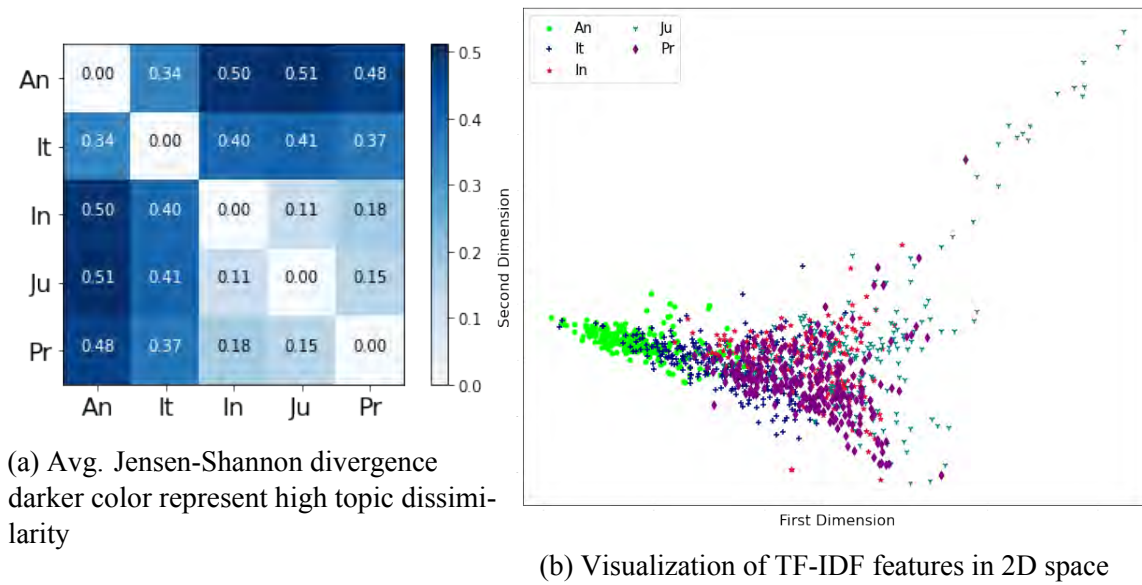


Figure 3.4: K-means clustering visualization for all authors using TF-IDF feature set *without-stopwords*

performs better since editorial articles are short, and removing the function words can decrease the available information like short stories in literature corpus. Similarly, stylometry feature set is not effective in the newspaper corpus since most of the articles are short and concise that might not represent the stylometry signature of the editors.

Table 3.11: Performance on Newspaper corpus

Criteria	Sample #	$w2v$ graph	$w2v+graph$ words	TF-IDF	Stylometry
An, It, Pr	765 (k=3)	.455 (s)	<b>.786 (c, m) (s)</b>	.762 (s)	.416
It, Pr, Ju	730 (k=3)	.572 (s)	<b>.619 (c, m) (s)</b>	.604 (s)	.517
Pr, In, Ju	678 (k=3)	<b>.691 (s)</b>	.526 (c, m) (s)	.440 (s)	.634
It, Pr, In, Ju	932 (k=4)	.515 (s)	.472 (c, m) (s)	<b>.554 (s)</b>	.505
An, Pr, In, Ju	935 (k=4)	.455 (s)	<b>.701 (c, m) (s)</b>	.553 (s)	.507
All 5	1189 (k=5)	<b>.478 (s)</b>	.391 (c, m) (s)	.382 (s)	.431

### 3.3.4 Feature Selection for Various Feature set

Since we utilized pre-defined features for both *Word2vec graph* and Stylometry feature set, we try to explore which features particularly contribute to k-means clustering for author identification or genre detection task. Therefore, we quantify the usefulness of these features that are defined by the features' discriminative power to tell clusters apart [105]. We apply the method in [106] to examine the means for each cluster on each dimension using analysis of variance (ANOVA) to assess how distinct the clusters are. The magnitude of the F values performed on each dimension is an implication of how well the respective dimension discriminates between clusters. We also validate these findings by sample chi-square test and tree-based feature selection method [107].

### Word2vec graph Feature-set

No. of *core* edges, *boundary* nodes, *core* nodes having *core* edges are the most significant features in differentiating between short stories and novels. Since the amount of text in short stories is limited, similarity among most frequent words is higher than other domains. The number of unique words is also less for short stories. Thus, it increases the number of *core* edges and reduces *boundary* nodes for short stories. Figure 3.5 shows the average value of some most important features in clustering on short stories and novels for various authors.

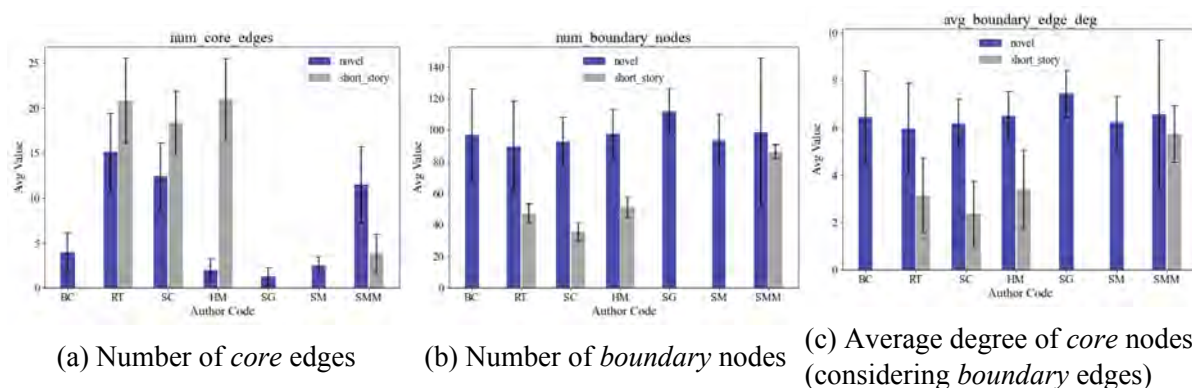


Figure 3.5: Average value (with standard deviation marked) of some features in *Word2vec graph* for novels and short stories of various authors.

No. of *core* edges plays an important feature in most of the tasks since it demonstrates the relationship between most frequent words of the story. Min/Max/Avg of *multiple/boundary* edge No./weight/degree also works as significant attributes for author identification in particular genres. For example, Humayun Ahmed science fictions show a higher number of *multiple* edges than Muhammed Zafar Iqbal, and the average index of them is also more. It indicates that *core* node associated words have higher similarity with words with lower index (more frequent) than words with higher index (less frequent). Degree of various core nodes also plays as a discriminative feature in author identification for long historical novels. We present some examples of sample features for author identification in different genres in Figure 3.6.

### TF-IDF Feature-set

Different word unigrams perform as discriminative features in separate tasks, and patterns vary for distinct authors and genres. In distinguishing between series novels of Humayun and Sunil, TF-IDF value of words associated with the protagonist and related characters play as the most important features. Some specific nouns (universe, revolver, detective, etc.) and verbs unigrams play as discriminative features for identifying specific genres, such as science fiction and thriller. Nouns and noun phrases are also important for separating historical novels from generic novels. The use of stopwords plays the most important role for author identification in short stories.

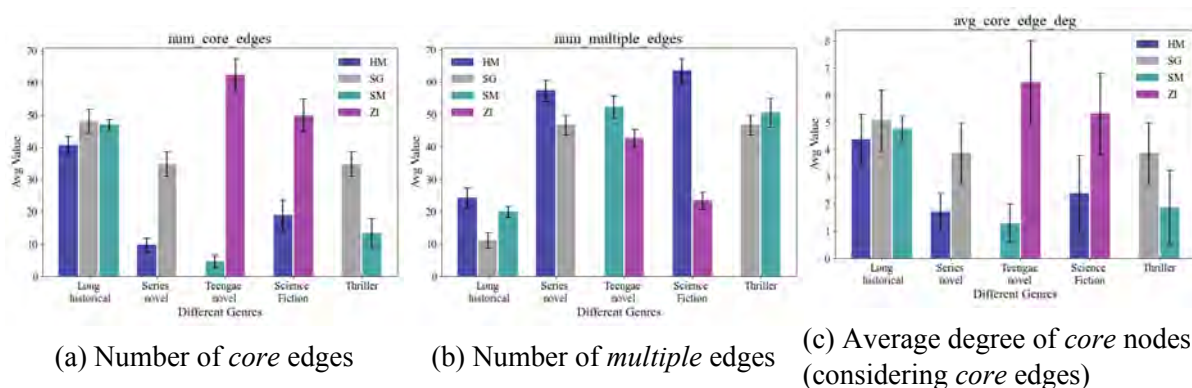


Figure 3.6: Average value (with standard deviation marked) of some features in *Word2vec graph* for different domains.

Author identification in novels mostly benefits from the choice of verbs (particularly *Sadhu* or *Cholito* form, also depends on the origin of the author) adjectives, and noun phrases.

### Stylometry Feature-set

Although Stylometry feature set does not perform equally as TF-IDF or *Word2vec graph* feature set in these tasks, some stylometry signatures become significant in the writings of various authors. Total sentences/words/characters is an explicit feature to differentiate short stories from other genres. It also becomes evident in some other criteria, such as teenage novels of Muhammed Zafar Iqbal are shorter than Shirshendu. Each author also follows an almost similar distribution of sentence length in all of his writings. Figure 3.7 shows the distribution and effect of sentence length in stories of various authors. Previous authors, such as Rabindranath, Bankim, Sarat mostly used larger sentences in their writings than contemporary authors, and the use of different lengths of sentences remains the same. Recent authors mostly use short sentences (the highest percentage is observed for sentence length four/five). Usually, usage of sentence length does not change much based on genre as we can observe the example of sentence length distribution of different genre writings for Humayun Ahmed. Interestingly, word length distribution shows a fixed pattern for all stories and does not change much based on genre/author.

Average stopwords per sentence works as an essential characteristic in most of the clustering tasks. Rabindranath and Sarat used more stopwords than other authors in novels and short stories. Vocabulary richness (No. of unique words/No. total words) is higher for short stories than other genres because of the limited text amount of short stories. Rabindranath and SaratChandra show higher vocabulary richness than modern-day authors. However, higher vocabulary does not mean better writing capability or author superiority. It is also an excellent qualification to express one's perspective efficiently using a limited number of unique words.

Sentiment or emotion word distribution does not perform as a significant feature for author identification in traditional novels. Still, it improves the clustering results in differentiating

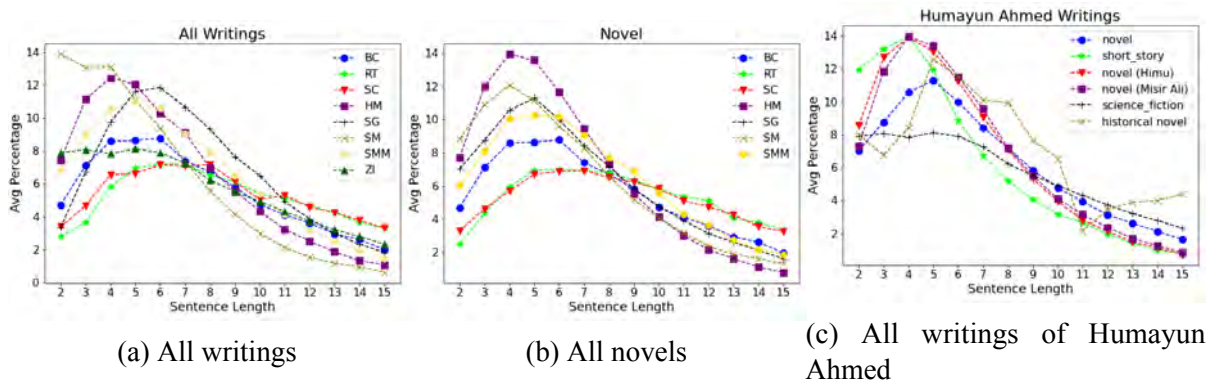


Figure 3.7: Distribution of sentence length in various writings.

teenage novels or thrillers from traditional novels since these genres show more fluctuation of sentiment or emotion. Although writings of all authors show an almost similar distribution of words starting with a specific character, it serves as a necessary feature in genre detection from the writings of any particular author. We show the impact of some sample features for various genres and authors in Figure 3.8.

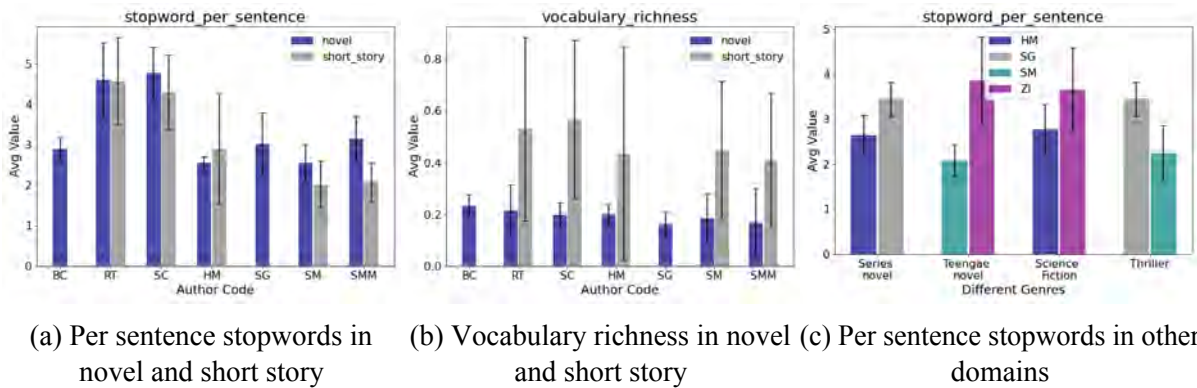


Figure 3.8: Distribution of various stylometry features.

Finally, we try to investigate whether stylometry features significantly change for an author in his writing career. From Table 3.8, it is evident that clustering performance based on the timing of writing using the Stylometry feature set is not promising. Moreover, only utilizing a subset of novels (whose publication date is available) is not sufficient to validate any hypothesis regarding the effect of stylometry feature change on their writing career. We provide some sample observations regarding the effect of stylometry features for contemporary authors over time in Figure 3.9. Notably, the average word length decreases gradually for previous authors, such as Bankim, Rabindranath, and Sarat. Their initial writings were inspired by the Sanskrit language that results in using larger words. They also show a decreased use of vowels in words over time. However, Humayun and Sunil show constant average word length. Vocabulary richness, average sentence length, stopword usage which are significant features for author identification, do not show much variation across writing years for all authors.

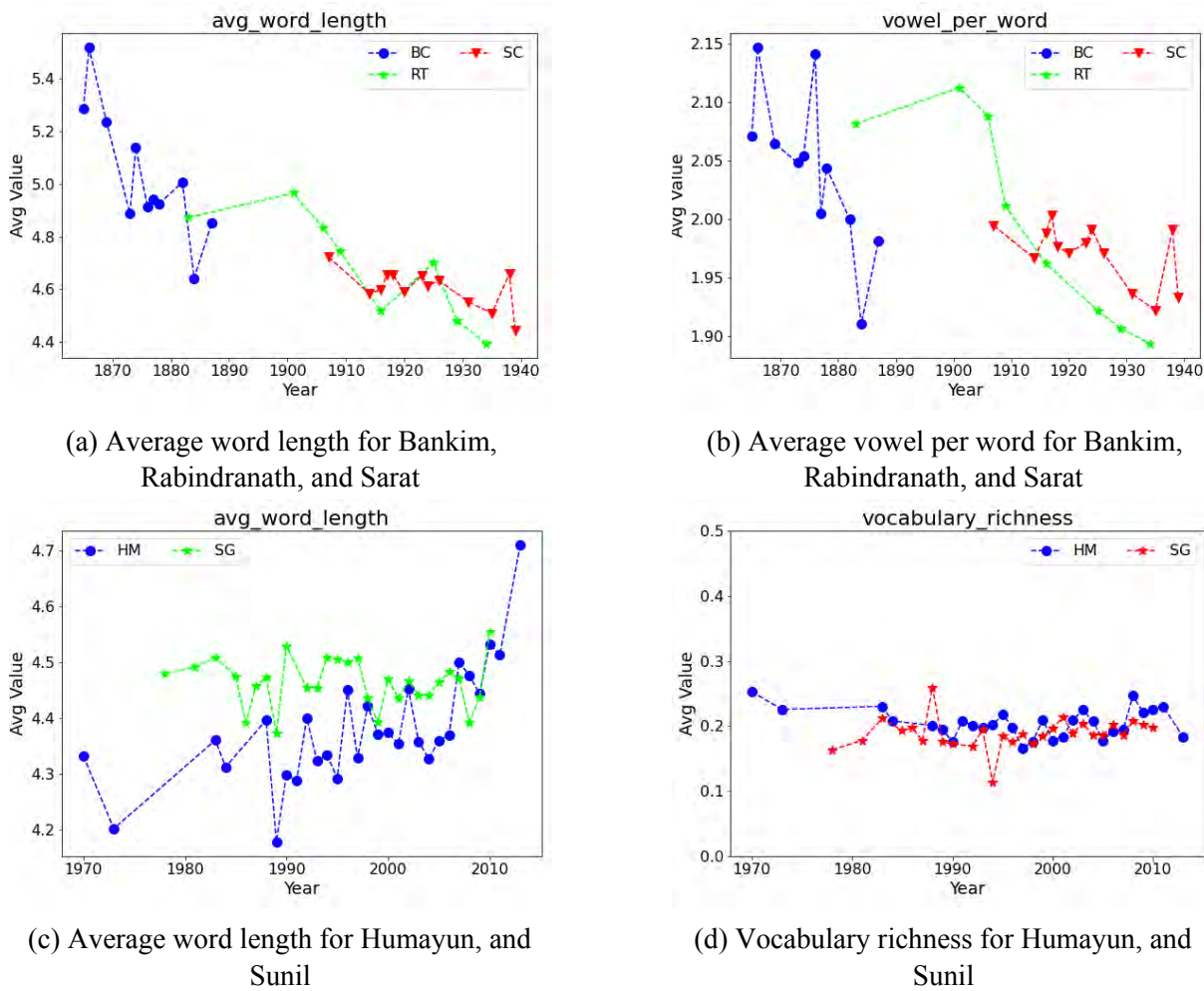


Figure 3.9: Change of different stylometry feature over writing years.

### 3.3.5 Discussion and Limitations

We compare our proposed *Word2vec graph* based approach with both bag of words (unigram) and various stylometry feature sets in author attribution, genre detection, and stylochroometry problems. Our approach consistently outperforms the baseline methods in various datasets irrespective of language and domain. Our model performs the best in genre detection tasks and achieves almost equal/ sometimes better performance than the TF-IDF feature set in the other tasks even though it has significantly fewer number selected attributes. We also show the effect of stopwords in various classification tasks. Along with solving stylometry tasks, we also identify the most prominent features from *Word2vec graph* and other two baselines that mostly contribute to various stylometry tasks. We discuss some interesting findings in our study in previous sections. Because of the limited amount of text in short stories, *Word2vec graph* can not properly capture their representation with its structure only. *No. of core edges*, & other related features play an important feature in most of the tasks since it demonstrates the relationship between the most frequent words of the story. Also, the stylometry feature set is not effective in



the newspaper corpus since most of the articles are short and concise that might not represent the stylometry signature of the editors. The most frequent words in each story of Project Gutenberg corpus are stopwords, and so they do not play any effective roles in clustering.

There are some limitations to our study. Since the major focus of our research is to introduce a new feature set, we employ a fairly simpler clustering scheme in our tasks. Also, the dataset size is relatively small because of the availability of the Bengali corpus. There, we would like to improve *Word2vec graph* by incorporating more variations and including most contributing features from other feature sets. Instead of handcrafted features, we will investigate how to graph embedding methods can be used to represent the graph and perform other classification tasks. We also want to extend the volume of our experiments in other corpora of different languages or domains. We believe that *Word2vec graph* could be an efficient approach to represent any text document. Utilizing words from *Word2vec graph* can be an alternative to the methods that employ all/most frequent words of the document in any NLP tasks.

It is evident that *Word2vec graph* can improve various literary analysis tasks in Bengali literature by building insightful graphs from story text. However, literary fiction adopts the form of narrative to report the events of telling a story through its plot and characters [108]. Utilizing the character aspects from fiction can provide some discerning details about the influence of contemporary society and social structure portrayed in stories. Therefore, in the next chapter, we employ character interaction graphs to answer these queries.

# Chapter 4

## Understanding Social Structures from Character Interaction Graph

We utilize character interaction graphs to answer a wide range of social questions regarding the influence of contemporary society on literary fiction. Our study involves constructing character interaction graphs from fiction, extracting graph features, and exploiting these features to resolve these queries. First, we describe our dataset and character interaction graph construction procedure. In the next section, We highlight our results and significant findings. Finally, we attempt to answer our research questions leveraging these findings.

### 4.1 Methodology and Experiments

We resort to character interaction graph model to answer our research questions (as outlined in the introduction) in the context of Bengali literature. Character networks of literary fiction have been discussed in several studies [70, 71, 108], where their generation procedure requires the assistance of many NLP tasks, such as Named Entity Recognition (NER), sentence tokenization, and others. Being an low-resourced language, such functionalities are not readily achievable in the Bengali language. Similarly, we can not adopt the techniques in [82] to construct character network for Bengali fiction as they are quite different from drama/play. Character names in drama are readily accessible, and the direct sequence of dialogues can detect the interaction easily. In other forms of fiction, these characters and interactions can not be identified easily. Besides, we are interested in explaining character networks in fiction from contemporary social structure, which requires some character and relational attributes that are not discussed in earlier studies for literary fiction. Therefore, our novel contributions in modeling and utilizing the character interaction graphs for these tasks are as follows.

- We adapt the procedure to construct the character interaction graph from story text and character list for Bengali fiction.

- We compute weight, sentiment score, and other attributes of nodes & edges from the story narrative and interaction.
- For our analysis, we also include some descriptive details such as age, gender, social group, religious characteristics for characters.
- We create a novel dataset containing the fiction of five prominent Bengali authors for character network analysis.

### 4.1.1 Character Interaction Graph Generation

Given a story, we construct the character network  $G := \langle V, E \rangle$ , where  $V$  is set of vertices, and  $E$  is the set of links between vertices. Each vertex  $v \in V$  is defined to be a character in the fiction, and each link  $e := (u, v) \in E$  is defined as the interaction between two characters  $u$  and  $v$ . Both nodes and edges contain several attributes in the character network. The process of extracting character interaction graphs from literary text mostly consists of three primary steps: 1) identification of characters, 2) detection of their interactions, and 3) extraction of the interaction graph [46]. Therefore, we have to modify these steps that would be applicable to our analysis in Bengali fiction. Since stories are collected in chapters, we perform these tasks and create character interaction graphs for each chapter like previous researches [67, 109]. Finally, we combine these chapter-wise graphs to construct the overall story graph.

#### Character Identification

Character identification consists of detecting which characters appear in the story and precisely when they appear in the narrative. Although some recent studies utilize NER to detect proper nouns from the story text and later perform unification in character occurrence to finalize the character list, this approach often misses characters mentioned in pronouns or nominal forms [46]. Also, current NER approaches [110, 111] in the Bengali language are not adequate to find correct character names in the context of literary fiction. In fact, identification of characters from story text using his proper noun forms, pronouns, and dialogues is itself a separate research problem. Therefore, we employ a manual direct annotation approach for character identification in our study.

For each story, we manually create a character list. The character list is verified from the synopsis of the story, by going over the narrative. If a character has multiple aliases, we include all the names it takes along with the usage of pronouns (stories described in the first-person narrative) in the character list. Hence it eliminates the necessity of unification of character occurrence step. We get the character list of a story verified by a reader who has read the story earlier. To find the occurrence of a character in the story text, we insert relevant suffix and inflection for each name, which is added after the noun form to make relation with the other words of the sentence. We

assume that if a character's name is appeared in any part of the story, he/she is present in that part. Otherwise, it would be required to manually validate each sentence whether a character is present or not, which would be troublesome. In our cases, character names can have two parts (first name and last name). Therefore, we consider both unigrams and bigrams while finding character occurrence.

### Interaction Detection

After character identification, we need to detect all interactions happening in the narrative between each pair of characters. Many researchers suggest that it is sufficient for a simple co-occurrence between two characters to infer an interaction between them [46]. In our study, we consider the sentence as the unit form of narrative. We assume that two characters interact when they appear in the same sentence or nearby sentences. Figure 4.1 demonstrates a sample instance of character interact identification procedure. We briefly discuss the interaction detection steps as follows.

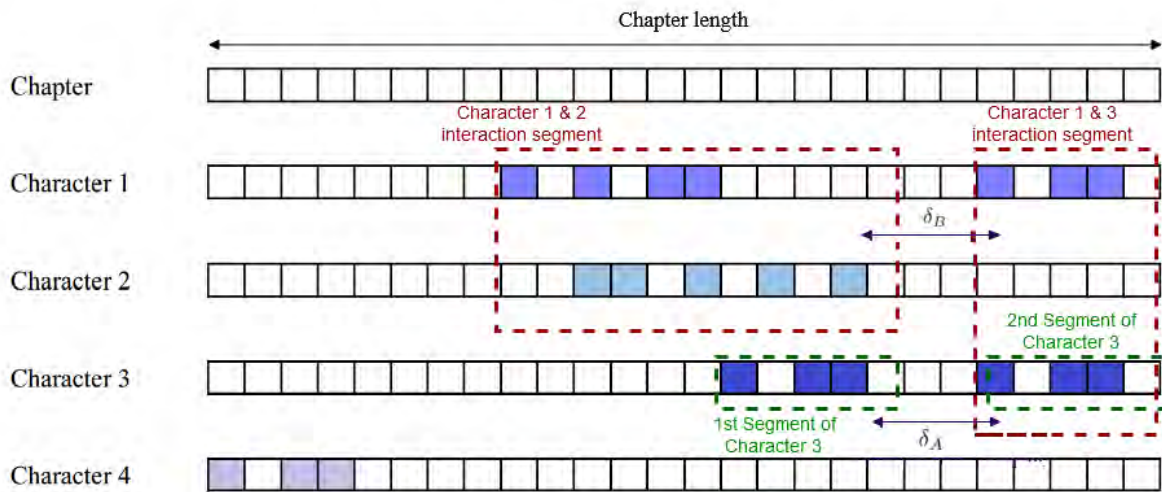


Figure 4.1: Interaction of different characters in a sample story chapter. Each sentence is denoted as a cell. A filled cell indicates that specific character is appeared in this sentence. Segment 1 and 2 of Character 3 are considered separate segments because their distance is greater than  $\delta_A$ . Similarly, segment 2 of character 1 and segment 3 of character 2 do not belong to the same plot since their distance is greater than  $\delta_B$ .

1. For each character  $C$ , we utilize the character occurrence to split the chapter into smaller segments. Now, two sentences  $i$  and  $j$ , where the character  $C$  is present, will belong to the same segment if  $|i - j| < \delta_A$ . Here,  $\delta_A$  denotes the threshold for intra-character sentence-wise distance. Also, a segment will form only if the number of sentences where the  $C$  is present is greater than the minimum appearance threshold value  $\delta_C$  in that chapter. Each segment will be denoted as  $S := \langle s, f \rangle$ , where  $s$  and  $f$  indicate the starting and finishing sentence of that segment.

2. Two characters  $C_1$  and  $C_2$  will have common a segment if any individual segment of  $C_1$  overlap with another individual segment of  $C_2$ . More specifically, characters  $C_1$  and  $C_2$  will have intersection if  $\langle s_1 - \delta_B, f_1 + \delta_B \rangle$  and  $\langle s_2 - \delta_B, f_2 + \delta_B \rangle$  overlaps, where  $\delta_B$  is the inter-character sentence-wise distance threshold.
3. We utilize these character and interaction segments to calculate weight, sentiment score and other attributes for nodes and edges in character interaction graph.

The values of  $\delta_A$ ,  $\delta_B$ ,  $\delta_C$  are automatically determined from the chapter length and characters count for each chapter of a story.

### Graph Generation

After character and interactions have been identified properly, we construct character interaction graph based on this information for each chapter and combine them to formulate the story graph. A character will be represented as a vertex if it is present in at least one segment in that chapter. A link will exist between two vertices if the corresponding characters interact in at least one segment. We compute node, edge weights based on their corresponding segment lengths, appearance, and other characteristics. We also include sentiment score, sequence, and supplementary information for nodes and edges. In all scenarios, we denote sentence-wise length as length. We briefly discuss these attributes in the following points.

- **Node weight:** A character's weight in a chapter segment depends both on the segment length and the number of times the character is addressed. Also, subsequent segments for a character in a chapter should indicate its higher weight than other characters that are present in fewer segments. Therefore we consider an incremental factor  $\alpha$  as the number of segments increases for a character.

Given a character  $C$  is present  $s_C$  segments in a chapter, length of the segment  $i$  is  $l_i$  and  $C$  is addressed in  $l'_i$  sentences in that segment. If the total chapter length is  $L$  and  $\beta$  is the extra weight for the sentences that contain character  $C$ , the weight of the corresponding node is defined as

$$\omega_C = \frac{1}{L} \sum_{i=1}^{s_C} (1 + i \times \alpha)(l_i + \beta \times l'_i)$$

- **Link weight:** Similar to node weight, we adopt a frequency-based [70] method to calculate edge weight. Interaction weight between two characters  $C_1, C_2$  depends on the number of segments they interact with  $s_{C_1 C_2}$ , segment length  $l_i$ , number of sentences they are present individually  $l'_i$  and number of sentences they are present both  $l''_i$ . The corresponding weight

of the edge is defined as.

$$\omega_{\langle C_1, C_2 \rangle} = \frac{1}{L} \sum_{i=1}^{s_{\langle C_1, C_2 \rangle}} (1 + i \times \alpha)(l_i + \beta \times l'_i + \gamma \times l''_i)$$

We assign 10% increment for subsequent presence of character and occurrence in same sentence. Therefore,  $\alpha = 0.1, \beta = 0.1, \gamma = 0.2$ , are considered in our study.

- **Sentiment score:** Sentiment of the character and interaction is derived from the segment they are present in the chapter. First, we calculate sentiment of each sentence in the chapter. For any segment, we interpolate the sentiment scores of the sentences in that segment to a fixed size to maintain uniformity. We calculate character (node) or interaction (edge) sentiment from the associated segments' average sentiment score.

Existing works in sentiment analysis in Bengali language only focus on specific domains, such as Twitter [93], YouTube comments [94], blog posts [95], etc. These approaches are not suitable to detect sentiment in literary text. Therefore, we only consider the sentiment of words in the document using Bengali SentiWordNet [96] and WordNet Affect [97]. However, there is a limited number of words in both of them, and they do not cover a significant portion of our corpus. So, we further improve them by correcting several entries manually and incorporating more words from English language resources, such as SenticNet4 [98], AFINN [99]. There are 9321 words in our version of SentiWordNet, and each has a score between -1 to +1. We utilize this SentiWordNet to calculate sentence sentiment as the average sentiment score of the tokens in the sentence. Although this method is reasonably simplistic, it can provide us an overview of sentiment associated with each character or relation.

- **Importance** We also calculate the importance of each character in their interaction as an edge attribute. Given two characters  $C_1, C_2$ , their segment length be  $l_1, l_2$  respectively and  $l$  is their overlapping length. Then the importance of character  $C_1, C_2$  in their corresponding link  $\langle C_1, C_2 \rangle$  is defined as.

$$\Phi_{C_1} = \frac{l}{l_1} + \frac{\# \text{ times } C_1 \text{ addressed}}{l}, \quad \Phi_{C_2} = \frac{l}{l_2} + \frac{\# \text{ times } C_2 \text{ addressed}}{l}$$

- **Other attributes** For each character and interaction, we also maintain their sequence (specific position of associated segments in the chapter), total appearance, and segment count.

**Story graph construction** We construct the total story graph from the weighted contributions of all chapter-wise graphs of that story. Node weight, link weight, sentiment score, character

importance all are computed from chapter graphs. We assume chapter weight proportionate to their length. Figure 3 depicts an example of a story graph from chapter graphs.

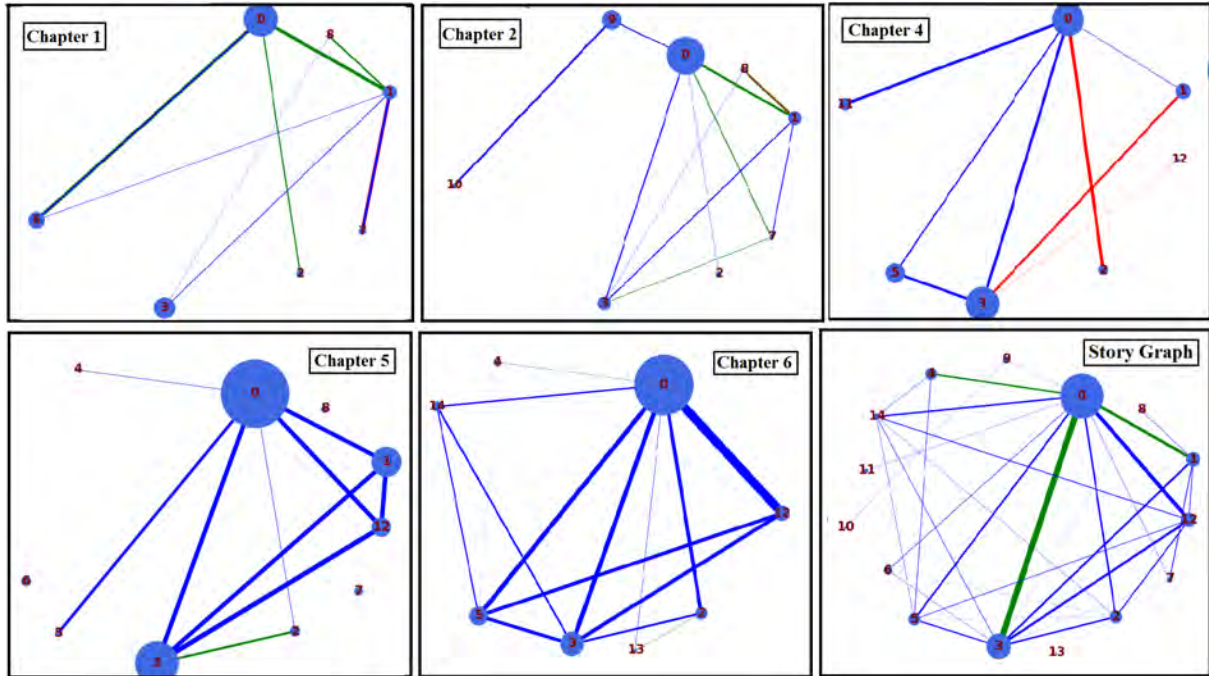


Figure 4.2: Generation of story graph for novel দেবী চৌধুরানী (*Devi Chowdhura,ni 1884*) by Bankim Chandra from corresponding chapter graphs. The final story graph include all nodes and edges that are present at any of the chapter-wise graphs. The node and edge weights in final are computed from the weighted average of these graphs. Blue, green and red edges indicate neutral, positive and negative sentiment respectively.

### 4.1.2 Graph Features Extraction

We extract different node, edge, and graph features from the static total graph of each story to answer our hypothesis. Although considering static graphs only may result in information loss, as they hide the chronology of the interactions [46]. However, we aim to explore the effect of contemporary social structure and events in the social interaction of corresponding fiction. And, features from the story graph could provide us sufficient information to investigate this interaction. Therefore, we do not consider chapter-wise graphs in this study and leave them as a scope for our future research direction.

#### Node Features

We consider weight, sentiment score, chapter presence for each node, which are readily available as attributions of the story graph. Also, we extract degree, strength (sum of weights over the edges attached to the node), and several structural information, such as closeness centrality [112], clustering coefficient [113], square clustering coefficient [114] for each node. Note that visual

inspection of story graphs does not demonstrate any particular community among vertices since fiction in early Bengali literature introduces relatively fewer characters, and their plot does not contain any specific division among characters [115, 116]. Therefore, we do not extract community detection related attributes from the graph.

Apart from the structural features of nodes, we also consider some descriptive information like age, gender, social group, religion features for corresponding characters since we attempt to leverage this information in light of contemporary social structure. Therefore, we have to manually annotate this information to the respective characters. We consider protagonist status (whether a character is protagonist/antagonist), gender (male/female), age group (discussed in next paragraph), family status (whether this character contains a family member role: father/mother/uncle/aunt/brother to the protagonist or central characters), religion, social status (poor/rich/landlord, etc.). Note that apart from gender and age not all attributes are available for all characters. We annotate these attributions from Wikipedia and critical analysis of the fiction. Along with the character list, we get verification about the annotation from an individual reader who has read the respective story and remembered the plot.

Although previous studies that discuss the role of different age groups in fiction and film consider various groups at ten years span [25, 48], age information for most of the characters is not directly mentioned in our stories. Therefore, we estimate three distinct age groups for our research purpose that show resemblance with real life. Throughout the rest of the paper, we utilize the term A1, A2, A3 to address these age groups.

- Age group A1: <20 year: This group mostly consists of children and adolescents. Since early marriage was regular in that contemporary era, members of this group can also show romantic relations with other characters.
- Age group A2: 20–40 year: Young adults and early middle-aged persons form this group. They are the majority portion of the story in most cases and serve as the current generation in story.
- Age group A3: >40 year: Older people. They usually play the role of the previous generation of young people (A2 group).

### Edge Features

Similar to nodes, we consider weight, sentiment score, common nodes, relative importance of two corresponding nodes, and other attributes for all edges, which are available in the story graph. Besides we extract edge\_betweenness\_centrality [112] and some other features from the graph structure. The descriptive features on edges could be found from corresponding vertex information.



## Graph Features

Different topological measures are used in the literature to describe character networks in various authors' writings [46]. We mostly consider graph density (ratio of its observed to possible edges), distribution of weights, and degrees of vertices and links as graph features.

### 4.1.3 Datasets

To understand the social structure from contemporary literature, we mostly focus on three prominent authors (Bankim Chandra (BC), Rabindranath Tagore (RT), Sarat Chandra Chattopadhyay (SC)) at the beginning period of modern Bengali literature. This period reflects a transitional colonial society with a rich history of its own. The elusiveness of social consciousness gives us a profound opportunity to research the role of contemporary society in literature. We consider their fiction, more specifically novels since works of fiction possess some specific features that are absent from non-fiction like essays [46]. Also, the length of short-stories is not sufficient enough to portray any proper character interaction. Since we want to make a comparative analysis between the authors, we do not consider dramas because Bankim Chandra and Sarat Chandra did not produce enough plays like Rabindranath. So, we consider novels only, and these novels belong to one or more genres, such as historical, romantic, social, and political. To make a resemblance with modern time writings, we also analyze several novels from two notable authors of recent period: Sunil Gangopadhyay (SG) and Humayun Ahmed (HM). We utilize a subset of our previous discussed corpus in Section 3.2. We extract character interaction graphs for a nearly equal number of novels for each author to keep uniformity. Throughout the rest of the paper, we use the first name or short form of the authors to represent them. A summary of the dataset is provided in Table 4.1.

Table 4.1: Overview of dataset for character interaction

Author	Career	# Collected Novels
Bankim Chandra (BC)	1865-1885	12
Rabindranath Tagore (RT)	1883-1935	11
Sarat Chandra Chattopadhyay (SC)	1907-1940	16
Humayun Ahmed (HM)	1970-2011	15
Sunil Gangopadhyay (SG)	1965-2012	14

## 4.2 Results and Findings

We analyze the extracted features from character interaction graphs for each author. Besides, we observe the trends in the change of specific features over time. We present our findings from different perspectives. First, we reveal the role of different age and gender groups by identifying

their presence, weight, and chronological changes. We also review the characteristics of protagonists and family members in different stories. Finally, we discuss whether the topological structure of character interaction graphs shows any difference in various fiction. We highlight our key findings as follows.

### 4.2.1 Age and Gender Distribution

How age and gender are depicted in popular media is an interesting area of study [47, 48] and can portray social structure and represent the authors' perspective [24, 117]. Table 4.2 demonstrates the proportion of different age & gender groups (as discussed in the previous section) and the mean (over stories) of aggregated weight  $\omega$  for specific groups in all authors. Similarly, average degree count of different age & gender groups is shown in Table 4.3. The presence of male characters is higher than female characters for all authors, which shows resemblance with the previous researches in other media [47, 48]. Also, the age group A2 is more represented than A1 & A3 for all authors. However, census reports in contemporary times show approximately equal male and female ratio (Table A.1 in Appendix). Similarly, the division of ages does not comply with the actual world distribution. Therefore, we observe a different distribution of age & gender in stories for the sake of plot.

Table 4.2: Age & Gender Proportion and Aggregate Weight for Each Group in Different Authors. All Values are Indicated in Normalised Form.

Author	Male	$\omega_{\langle M \rangle}$	Female	$\omega_{\langle F \rangle}$ (%)	A1	$\omega_{\langle A1 \rangle}$	A2	$\omega_{\langle A2 \rangle}$	A3	$\omega_{\langle A3 \rangle}$
BC	0.6086	0.4273	0.3913	0.5727	0.0729	0.103	0.5620	0.7327	0.3649	0.1643
RT	0.6410	0.5773	0.3589	0.4227	0.0172	0.0293	0.6982	0.8497	0.2844	0.121
SC	0.6936	0.5457	0.3063	0.4543	0.0990	0.103	0.6081	0.7327	0.2927	0.1643
HM	0.6726	0.6669	0.3273	0.3331	0.0778	0.0467	0.5628	0.6234	0.3592	0.3299
SG	0.7058	0.7573	0.2941	0.2427	0.0441	0.0292	0.7794	0.9047	0.1764	0.0661

Bankim Chandra and Humayun Ahmed have more aged characters (A3 group) in their fiction. Bankim Chandra's novels contain an atmosphere that is surcharged with an aroma of the feudalistic structure of society [40] that incorporates more A3 characters. Humayun Ahmed mostly wrote about the struggles of middle-class families in contemporary times [118] and his fiction thereby also includes a higher number of older characters. Sarat Chandra has a similar presence of A1 character group in his fiction like Bankim. Both of their fiction include A1 characters that often play protagonist or central roles since early marriage of girls were very common in their background (rural society for Sarat and feudalistic society for Bankim [40]). However, the appearance of A1 group is their contemporary author Rabindranath is very limited since his fiction mostly includes characters coming from the higher stratum of the society, usually the upper-middle class, with its members have been the recipient of an enlightened liberal education [73]. Therefore, the presence of A2 group is also higher for Rabindranath than Bankim

and Sarat. We observe that the female percentage does not increase for the fiction of modern days authors, which shows an exception from existing studies in other media [47].

Table 4.3: Average Degree Count of Different Age and Gender Group.

Author	Male	Female	A1	A2	A3
Bankim Chandra	4.1441	<b>5.736</b>	<b>6.1667</b>	5.0067	4.8264
Rabindranath Tagore	<b>4.6456</b>	3.9335	4	<b>4.255</b>	3.8685
Sarat Chandra Chattopadhyay	3.7407	<b>5.7594</b>	4.6881	<b>5.3089</b>	4.2225
Humayun Ahmed	<b>5.102</b>	4.8974	<b>5</b>	4.193	4.1214
Sunil Gangopadhyay	4.7381	<b>5.5858</b>	4.1667	<b>5.3199</b>	3.9673

One interesting observation is that the overall weight is higher for female characters compared to their lower representation in the novels of previous authors. Critical analysis of stories also shows that most of the plots at the beginning of modern Bengali literature evolve around women from different aspects of life [38, 39]. However, for recent authors, gender-wise total weight is proportionate to their presence since their fiction do not provide any special priority to female characters. The older age group receive considerably less weight compared to their presence in most authors except Humayun, which can assert the background of his fiction (middle-class struggles in modern time with a significant presence of older characters) [118]. A3 age group also has the least connectivity in all authors (lowest degree count). Urban-centric plot allows male characters in Rabindranath to have more connectivity than female. Also, A2 group has the highest degree count for both Rabindranath and Sarat in previous times because of their more contemporary plots.

Table 4.4: Age and Gender Wise Combined Distribution for All Authors. All Values Are Indicated in Percentage (%).

Author	M-A1	F-A1	M-A2	F-A2	M-A3	F-A3
Bankim Chandra	2.4096	14.8148	49.3976	66.6667	48.1928	18.5185
Rabindranath Tagore	1.3333	2.439	68.0	73.1707	30.6667	24.3902
Sarat Chandra Chattopadhyay	7.1429	16.1765	61.6883	58.8235	31.1688	25.0
Humayun Ahmed	1.7857	20.0	56.25	56.3636	41.9643	23.6364
Sunil Gangopadhyay	4.4944	8.0	79.5506	88.0	15.9551	4.0

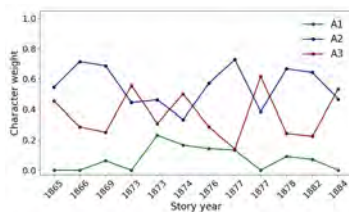
Table 4.4 shows the age & gender-wise combined distribution for all authors. We observe that A3 age group contains more males than females and vice versa for A1 group. Therefore, males in A1 age group are represented as children in most stories, where females in that group can show different roles (lover, widow, married, and others) that assert their higher proportion. Similarly, A3 age group mostly functions as the supporting characters, parents, or moral representation of society that incorporates more male characters due to the patriarch society of contemporary times. Bankim has the almost same ratio of A2 and A3 age group for male characters. Because

Bankim’s novels mostly incorporate a feudal society background, where kings, landlords, and other influential characters belong to A3 class. All three previous authors have nearly non-existent male characters as A1 age group that implies that children do not receive much attention in the social structure of that period. Female characters in A1 group receive attention because of their roles other than as children. Contemporary authors also show a similar pattern where presence of male children is significantly lower. It demonstrates that male children have always received less attention in Bengali literary fiction.

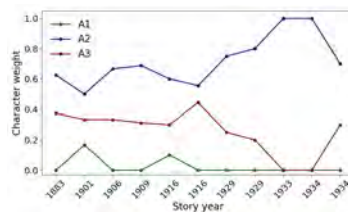
Table 4.5: Age-wise edge distribution in different authors

Author	M-M	M-F	F-F	A1-A1	A1-A2	A1-A3	A2-A2	A2-A3	A3-A3
BC	0.3423	<b>0.4878</b>	0.1698	0.0084	0.1204	0.0532	0.3053	<b>0.4006</b>	0.1204
RT	0.4328	<b>0.4477</b>	0.1193	0	0.0192	0.0115	<b>0.4981</b>	0.3793	0.092
SC	0.4216	<b>0.4612</b>	0.1170	0.014	0.1012	0.0471	<b>0.4887</b>	0.2914	0.0716
HM	0.3561	<b>0.4928</b>	0.1509	0.0088	0.0936	0.0643	0.3216	<b>0.4152</b>	0.1053
SG	<b>0.4984</b>	0.4264	0.0750	0	0.0511	0.015	<b>0.6727</b>	0.2282	0.033

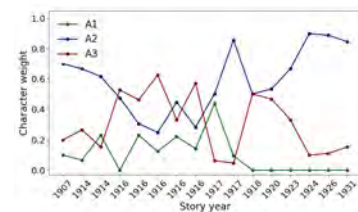
We also observe the interaction between different gender and age group in Table 4.5. The proportion of male-female link is highest for most authors except Sunil since his plots lack female characters and they have a more cornered role. For previous authors, Rabindranath and Sarat follow a nearly same distribution of gender-wise links. Also, the ratio of A2-A2 (the edge between two middle-aged/young characters) is highest in the age-wise distribution for them, where A2-A3 is the most present edge type for Bankim. Rabindranath’s urban-centric plot and Sarat’s rural society-based plot both allocate male and female A2 group as central characters and they do not interact much with the A3 age group as in the feudal social structure of Bankim’s novels. Humayun Ahmed’s plots also evolve around contemporary social life, family struggle, and therefore also demonstrates A2-A3 type as the highest edge group. Relation among children is very rare for all authors since none of these are children specific fiction. Also, Bankim and Sarat show a higher A1-A2 ratio than Rabindranath since many of their A1 characters actively participate in the story plot.



(a) Bankim Chandra



(b) Rabindranath Tagore



(c) Sarat Chandra Chattopadhyay

Figure 4.3: Distribution of different age group over time for different authors.

In our study, we are also interested to explore whether events in the real-world influence characters in fiction specifically at the beginning of modern Bengali literature. Therefore, we present the

ratio of different age groups over time for three previous authors in Figure 4.3. We observe that proportion of A1 age group increases during the 1873-1877 period in Bankim's writings after the approval of the widow remarriage law in 1872 (Table 4.9 in the next section). However, A2 and A3 age group do not exhibit any significant pattern for Bankim over time. A3 age group is more prominent in Sarat's fiction in 1916 during the economic crisis in the sub-continent due to World War 1 (Table 4.9) and much of Sarat's plots in that time revolve around the struggle in rural society. Moreover, we identify that the presence of A2 age group increases significantly from around 1918 and alternatively decreases for A3 age group in the writings of both Rabindranath and Sarat. Non-cooperation and other nationalist movements shift the plot of their fiction from romantic or conventional social issues to more political or social crisis-oriented plots that incorporates more A2 characters. Also, A1 group becomes almost absent in their writings during that period. We do not identify any specific patterns in the change of gender ratio over time. In the next section, we explain these phenomena in a more detailed way with relevant case studies according to our research questions.

#### 4.2.2 Protagonist Characteristics

A protagonist is a central character or leading figure in fiction who advances the narrative of the story through his/her action. Protagonists typically experience some alteration that triggers a turn of events in the plot. which makes the story compelling and provides the author's perspective [119]. Therefore, protagonists and their characteristics can deliver significant insights into the social structure portrayed in the novels. First, we provide the list of stories, genre details, and corresponding protagonist age and gender information in Table 4.8 for three previous authors. We confirm the story genre using the critical analysis of the authors, synopsis of stories, and with the help of Wikipedia and Goodreads. Note that, some novels can be categorized as multiple genres. Also, fiction in Bengali literature does not contain traditional antagonists like other forms of media, such as films.

Almost all protagonists are young (A2 class) in these stories. Bankim has some A1 female protagonists after widow remarriage law, and Sarat has one romantic novel with a female A1 group protagonist. Both Bankim and Sarat has one social novel with A3 group protagonist. Rabindranath represents all of his protagonists as A2 age group: young adults, later adulthoods, or middle-aged due to urban-centric plots. The gender of protagonists does not vary much according to the genre. Although all historical/romantic novels of Bankim have female protagonists, both Rabindranath (*Shehser Kabita*) and Sarat (*Devdas*) have romantic novels with a male protagonist. We observe a significant time difference in Rabindranath's novels since he took a break in fiction writing because of his world visit after the noble prize-winning. His previous social and political novels in the early twentieth century were based on male protagonists, where his later works in these genres include female leads. His inspiration could be drawn from the active involvement of

women in different nationalist movements in the 1920s [38, 80]. His contemporary author Sarat did not follow a similar trend as his social and political novels were still male protagonists based in the later 1920s. Instead, between 1916-1918, most of his novels had female leads, which are based on various social problems in the rural community. For both male and female protagonists, we observe similar importance in the interactions they maintain.

Table 4.6: Male and Female Protagonist Characteristics for Authors

Author	M (%)	F (%)	$\omega_{\langle M \rangle}$	$\omega_{\langle F \rangle}$	$\bar{D}_{\langle M \rangle}$	$\bar{D}_{\langle F \rangle}$	$\bar{S}_{\langle M \rangle}$	$\bar{S}_{\langle F \rangle}$
Bankim Chandra	0.6153	0.3847	0.5917	0.8724	9.22	9	-0.2384	-0.4113
Rabindranath Tagore	0.444	0.556	0.244	0.8974	9.2	7.85	-0.1482	0.116
Sarat Chandra Chattopadhyay	0.6	0.4	0.5111	0.5091	8	4.4	-0.1833	-0.2109

We also include the protagonist's average weight ( $\omega$ ), degree ( $\bar{D}$ ), and sentiment score ( $\bar{S}$ ) based on gender in Table 4.6. Because of his urban-centric plot and inclusion of women as the lead roles in political and social novels, Rabindranath has a higher female protagonist rate. In addition, his female protagonist's average weight is considerably higher than the male counterpart. For Bankim, the weight of the female protagonist is also higher than that of the male, but both genders in Sarat's novels have almost equal weight. Male and Female protagonists in Rabindranath and Bankim have nearly similar degree connectivity. However, Sarat's novels with female protagonists have less character due to a concise plot that explains its lower degree count for female leads. Finally, female protagonists are also more sentimental (both positive & negative) than males for all authors. We also observe in our study that the average sentiment of all characters and relations are slightly negative which contradicts with Pollyanna effect [120].

### 4.2.3 Influence of Family

Family plays a significant part in the context of the Bengali social structure [121]. We intend to examine how, during the era we are studying, the significance of family varies over time in Bengali literature. Therefore, we plot the story-wise average for the aggregated weight of characters with distinct family attributes over time in Figure 4.4 for all previous authors. In Bankim's writings, family involvement was infrequent since many of his novels follow the background of feudal society. It creates a social network of related individuals who are not family members of the protagonist, revolving around the landlord's character. In the writings of both Rabindranath and Sarat, up to 1916, we observe a greater family weight. After that, in their writings, we show hardly any family presence, especially those novels associated with contemporary social and political problems.

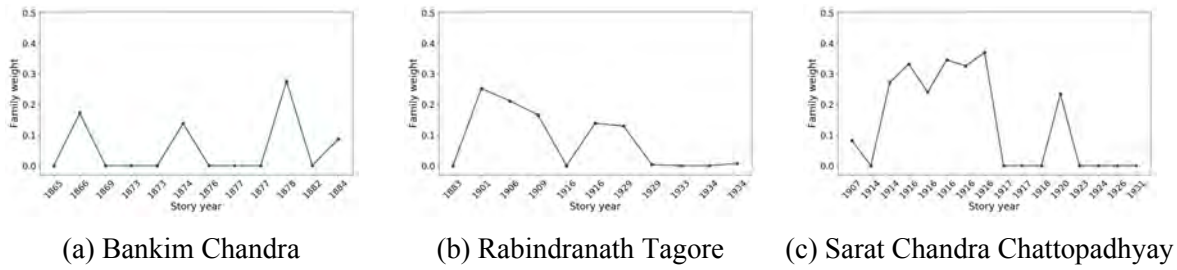


Figure 4.4: Weight associated with family members for different authors over time.

#### 4.2.4 Variation in Graph Structures

We also analyze whether, in different genres and authors, the graph structure demonstrates any substantial difference. We examine various structural properties of character interaction graphs for this intent. First, we present the average count of node, edge and graph density for all authors in Table 4.7. It shows that Rabindranath has a less number of nodes and edges. Sarat has more characters in his fiction but usually depicts smaller connectivity. The graph density is highest for Bakim. Since Rabindranath wrote about higher middle class educated urban society, it often involves less but compact social structure. On the contrary, the rural society background of Sarat involves more characters. However, we can not assert any tangible structural difference based on authors only.

Table 4.7: Average Count of Nodes, Edges and Graph Density for Different Authors

Author	# Node	Density	# Edge
Bankim Chandra	11.6923	0.4952	32.2308
Rabindranath Tagore	10	0.4565	19
Sarat Chandra Chattopadhyay	14.4375	0.363	37.25

For various genres, Figure 4.5a shows the graph density versus node count plot. Although no clear distinction between different genres is identified, we observe some interesting properties. Romantic novels, for instance, contain either a small dense network (less node count, higher density, such as *Yuglanguriya*, *Parinita*) or a large sparse network (higher node count, less density, such as *Sheher Kabita*, *Rajani*). Similarly, a substantial number of characters are typically included in historical fiction. Also, political novels preserve the high value of graph density even though the character count increases.

Node weight and degree distribution display minimal difference since they exhibit exponentially and linearly decreasing patterns for most of the stories, respectively. Visualization of the story graphs, however, may reveal some specific distinctions between them. Romantic novels, for example, generally include an edge (the edge between the central male and female character) whose weight is significantly greater than the other edges. Figure 4.5b represents the edge weight distribution for various genres. For each story, we normalize the edge weights with respect to

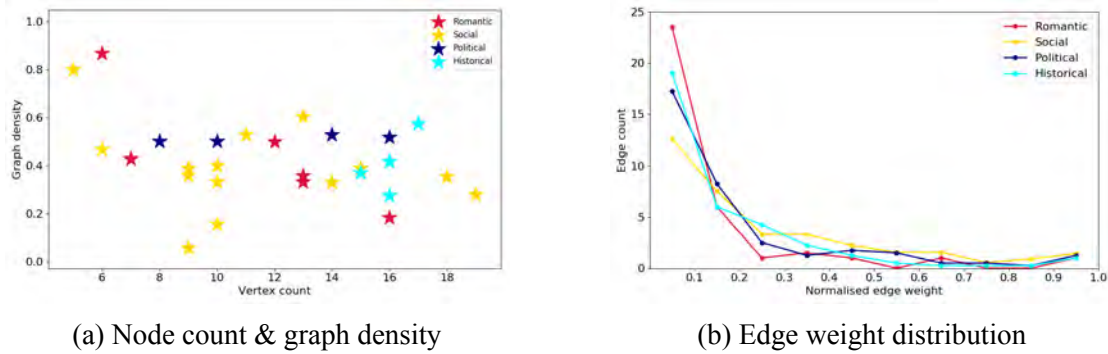


Figure 4.5: Graph topological properties for different genres.

the maximum edge weight of that story and linearly interpolate the edge weight list to a fixed size. Then we take the average of edge weights for all stories in that genre (if a story belongs to several genres, it is considered in both) and divide the edges according to their weight into ten partitions. Here, partition  $i$  contains edges whose weights are between  $[(i - 1) \times 10\%, i \times 10\%]$  range of the maximum weight. For all genres, the edge distribution in these partitions follows an exponentially decreasing distribution. The last bin contains only one edge, so we can assume that each story's most significant link has at least 10 percent more weight than the next one. Romantic and historical novels also have the most edges in the first partition. It means several irrelevant and passive connections are present in these genres that are not important to the plot. Besides, for these genres, the most significant link is almost double the weight of the next link, which is present in either the fifth or the sixth partition. There is an edge in nearly all partitions for social and political novels, suggesting the presence of various levels of interaction in these novels. Also, the decreasing rate is the lowest for social fiction.

## 4.3 Discussion

Based on our key findings in the previous section, in this section we answer three research questions. We validate these assumptions with the help of critical analyses from various perspectives in Bengali literature. We also discuss our limitations and the future scope of this study.

### 4.3.1 Influence of Real-life Events in Story Character Arc

To address whether historical or social events impact contemporary fiction, we first create a list of significant national events during our period of study in Table 4.9. We observe that the law of widow remarriage in Hindu society (1972) influences Bankim's contemporary novels. Besides, in the nineteenth century, multiple nationalist movements inspired Rabindranath and Sarat to produce several social and political novels. As a case study, we use the character interaction

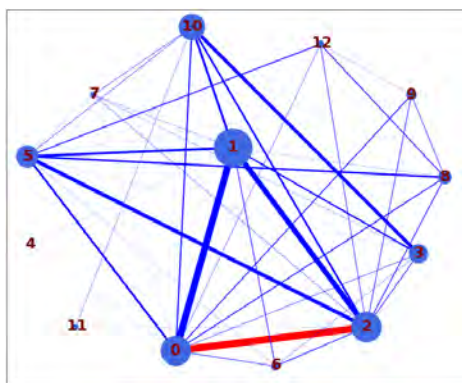


graph from corresponding novels to address whether these events induce an effect on the graph, and it can represent the contemporary social structure.

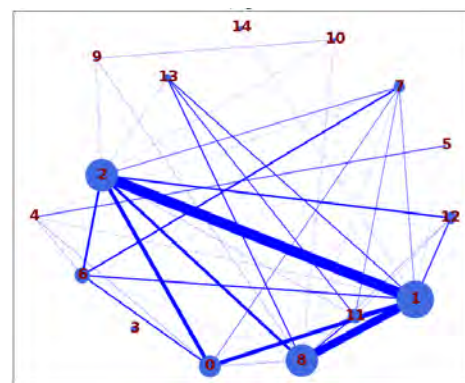
### Impact of Widow Remarriage Law

The Brahma Marriage Act passed in 1972 lifted the ban on widow remarriage. It is epitomized as the interplay of competing feelings and beliefs that characterized men's knowledge of women during this seminal period of Indian history [122]. Bankim had widows as his central concern, as he also wrote about widows in a number of his non-fictional writings and addressed the issue of their remarriage [116]. After the remarriage act, we observe a significant increase in the female A1 character category in Bankim's writings. Widows are the central characters in his two novels, *বিষবৃক্ষ* (*Bishabriksha*, 1872, Figure 4.6a) and *কৃষ্ণকান্তের উইল* (*The Will of Krishnakanta*, 1878, 4.6b).

**বিষবৃক্ষ (Bishabriksha, 1872):** The major interactions that happen in this story are between the protagonist Nagendra (Index 0, M-A2), one male character Debendra (Index 3, M-A2), four other female characters where three are from A1 age groups, and two are a widow. While the novel's protagonist is Nagendra, the widow character Kundra (Index 1, F-A1) has a higher weight in the story. The inclusion of this widow character creates a triangle of relationship between Nagendra, his wife Surjamukhi (Index 2, F-A2). Interestingly, the weight of Nagendra-Kundra relationship is greater and more positive than the relationship with his wife. Also, the other widow character Hira (Index 10, F-A1) receives significant attention and develops interaction with different male characters in the story. The protagonist Nagendra shows 92% degree connectivity, which conforms to his role as a landlord in contemporary society.



(a) *বিষবৃক্ষ* (*Bishabriksha*, 1872)



(b) *কৃষ্ণকান্তের উইল* (*The Will of Krishnakanta*, 1878)

Figure 4.6: Character interaction graph for two novels of Bankim. Protagonist (Index 0) of both stories are widow.

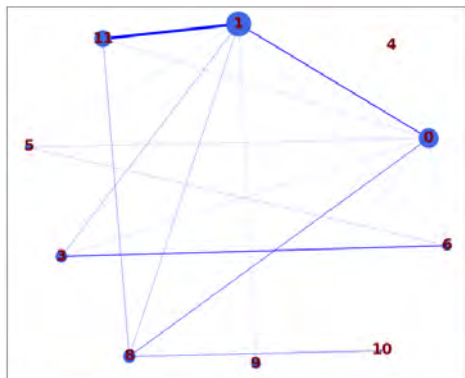
**কৃষ্ণকান্তের উইল (Krishnakanter Will, 1878):** Widow characters are represented in this social fiction as well. Here the central widow character Rohini (Index 2, F-A2) also serves as the other

woman in the triangle with Gobindalal (Index 1, M-A2) and his wife Bhramar (Index 8, F-A1). The weight of the relationship between Gobindalal and Rohini is slightly greater than that of his wife, Bhramar. Gobindalal demonstrates less degree of connectivity because of not being a landlord.

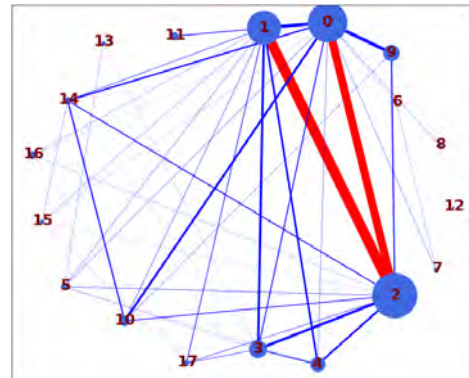
Bankim shows opposition to widow remarriage law by including widows in his novels to establish a relationship triangle with a married male and his lawful wife. The character interaction graph of these stories shows this relationship triangle as visually noticeable edges of the graphs and also portrays an overall negative sentiment. It is also confirmed by prior researches that the plots of these stories mostly lead to a negative ending [116, 122]. Two novels of Bankim during this period, যুগলঙ্গীও (Yugalanguriya, 1874) and রাধারানী (Radharani, 1876) both have female A1 characters as protagonists who are not a widow. Their character interaction graphs do not portray such a relationship triangle. Therefore, it is noticeable that Bankim illustrates his resistance to widow remarriage in his fiction by establishing a unique character arc.

### Impact of Nationalist Movement

In the first few decades of the nineteenth century, various nationalist movements influenced Rabindranath and Sarat to portray their perceptions through fiction. To understand whether it is possible to depict contemporary social structures, we briefly review the character interaction graphs of two political novels: ঘরে বাইরে (Home and Abroad, 1916) by Rabindranath (Figure 4.7a) and পথের দাবী (Pather Dabi, 1926) by Sarat fiction (Figure 4.7b), which are inspired by non-cooperation movements and other ongoing activities in Indian sub-continent.



(a) ঘরে বাইরে (Home and Abroad, 1916)



(b) পথের দাবী (Pather Dabi, 1926)

Figure 4.7: Character interaction graph for two political novels of Rabindranath and Sarat influenced by nationalist movement.

**ঘরে বাইরে (Home and Abroad, 1916):** The plot includes three primary characters, Bimla as the protagonist (Index 0, F-A2), her husband Nikhil (Index 3, M-A2), and a nationalist character Sandeep (Index 1, M-A2). Although Nikhil is described as the hero of the novel by critical

analysis [115], he received less attention in graph. Bimla also does not show the highest node weight, which is achieved by Sandeep. However, Bimla has a higher degree count indicating her connecting with more characters (most of them have some family attributes). Although Bimla was not a part of the nationalist movement, the relationship weight between herself and Sandeep is higher than her relationship with Nikhil. Therefore, this novel represents a small part of contemporary society where a typical female has to maintain a lot of relations with her family members but leans to an iconic figure rather than her husband. Also, the relatively small weight of the female protagonist can be explained by her non-participation in these movements. Later books of Rabindranath portrays higher weight for female protagonists in social/political novels when female participation in national movements was more common [80].

**পথের দাবী (Pather Dabi, 1926):** The plot revolves around a fictional order called "Pather Dabi" and involved primary characters are the protagonist Apurba (Index 0, M-A2), leader of the group Sabyasachi (Index 1, M-A2), Bharati (Index 2, F-A2), Sumitra (Index 3, F-A2), Tewari (Index 9, M-A2). Except for Sumitra, all major characters are connected to the organization. The negative emotion between Apurba-Bharati and Sabyasachi-Bharati indicates the tension between the group since they are not romantically involved [39]. Although not the leader, protagonist Apurba has more degree count than Sabyasachi. It is also supported by the plot of the story that Apurba later worked as the spy for the police. Several minor characters are a part of the order and form a relationship between them. Therefore, the character interaction graph shows the overall structure of a fictional nationalist organization of contemporary times.

Only one character in Ghare Baire was involved in the nationalist movement, where many characters in Pather Dabi are related to the organization. As a result, Ghare Baire shows less node count as well as many trivial edges with less weight. The node count, graph density, and overall edge weights are higher in Pather Dabi because of its context. Therefore, certain historical events have always inspired authors of contemporary times to represent their ideas through fiction. However, we can not confirm that corresponding social structures will always be reflected in this fiction. Rather they depict a very small part of the society, which may include some specific types of characters inspired by real events as we observed in our case studies. How these characters interact in the fictional social network, however, does not always depend on the precise social structure. It depends, instead, on the perception of these events by the authors and their imagination.

### 4.3.2 Influence of Different Age and Gender Group

We analyze the appearance of different age and gender groups and how they evolve for various authors at the beginning of modern Bengali literature. Our findings indicate that literary fiction presents a distorted picture of gender and age that does not comply with the real-world distribution.

In the writings of previous times, while the ratio of female characters is lower, they display a greater weight than males. Recent writers, however, do not always exhibit such increased weight for female characters. Fictions of different genres at the beginning of modern Bengali literature, thereby, include plots that revolve around women and their relations with society. Although political novels do not have female characters in major roles, the situation begins to shift with the inclusion of women in various movements.

Most characters are represented as the 20s-40s (A2 age group) for all authors. Many female characters in their 10s-20s (A1 age group) play as central characters because of their various roles. It definitely indicates the contemporary social structure that includes early marriage and the maturation process of females at an immature age. Events in real-life often impact the presence of different age groups in contemporary fiction. The older age group was prominent in fiction from the beginning of modern Bengali literature until 1916. After that, various social and nationalist movements create a paradigm shift, which increases A2 group participation significantly and almost makes other groups non-existent. The impact of family in fiction was not significant in nineteenth-century fiction. It becomes significant in social novels from the early twentieth century for nearly two decades. The influence of nationalist movements subsequently makes the fiction socio-political driven, which makes the involvement of the family negligible.

### 4.3.3 Interpretation of Character Interaction Graph from Context and Genre

Finally, we evaluate whether the topological structure of the graph or the presence of different characters can be inferred from fiction's background or genre information. For example, Bankim's novels provide an aroma of feudalistic social structure with landlords, kings as characters. Therefore, his fiction incorporates relatively more aged characters (A3 age group) than two other previous authors. Similarly, the urban-centric social plot in Rabindranath's fiction that evolves on upper-middle class educated characters, include more young characters (A2 age group than others). This background often provides more weight to the female protagonists than other authors even though they have less average connectivity. Also, his fiction does not include noticeable female A1 group characters like Bankim or Sarat. The rural social background in Sarat's novel often incorporates fewer female characters in general but their presence, connectivity and weight are more noteworthy. Apart from the novels: *Bouthakuranir Hat* and *Gora*, the node count and edge weights of minor characters in Rabindranath's fiction is significantly lower than Sarat. Therefore, urban-centric background often introduces fewer characters and less interaction than rural centered plot.

Sometimes, the fiction genre influences the structure of the graph and the appearance of characters. Romantic and political novels, for instance, mostly have female and male protagonists respectively. However, Rabindranath shows both male and female protagonists in these genres.

Romantic novels either contain a small densely connected network or a larger network with lower connectivity. It is also possible to visually distinguish story graphs of romantic novels with the presence of a thicker edge indicating the relationship between the central male and female character whose weight is substantially higher than other edges. There is a significantly higher graph density in political novels, which does not differ much as the number of nodes increases. Likewise, graphs of historical novels typically include a greater number of nodes. All these observations suggest that fiction's context and genre frequently impact the presence of the character and their interaction.

#### 4.3.4 Limitations

There are several limitations of our study. In this study, we focus on the influential Bengali writings of the late nineteenth and early twentieth-century social transitions. In this time range, we consider almost all novels of the three most influential authors of that period whose writings can provide us substantial material to conduct our study. As a low resource language, we keep our dataset limited to five authors only because the manual annotation of characters and their attributes requires enormous effort and detailed knowledge of these 58 stories. The manual listing of characters may create some unwanted errors, where some characters could be missing in the character interaction graph. However, these are minor characters, and they have a very marginal presence in the story context. Therefore, their absence should not impact the character interaction in graphs [46] and would not be able to compromise the quality of our analysis.

As a starting point of the research in this domain, we adopt a simple SentiWordNet based approach to calculate the associated sentiment of characters and interaction since existing sentiment analysis techniques in Bengali NLP are not proven yet in the literature domain. However, this approach is sufficient to provide us an overall sentiment vibe in the story context. Also, using a manually created list of characters for character identification may miss their presence when they are mentioned in pronoun or other forms. Finally, we only utilize the static graph (story-wise graph) in our analysis since we are particularly interested to observe the effect of contemporary events and the presence of different character groups in fiction. The change of interaction, sentiment, weight of different characters along with the storyline progression could be considered as a separate future study.

There are various unexplored character attributes in our study, such as religion and economic status, that we would like to consider to provide such analysis from a different perspective. Moreover, we aim to utilize dynamic graphs (chapter-wise graphs) so that it can identify the evolution of interaction between different characters throughout the story context. Finally, a more sophisticated approach for character recognition, their descriptive attribution identification, and sentiment/emotion detection from story text would be beneficial in future research.

Table 4.8: Year-wise Story, Genre and Protagonist Information for All Authors

Bankim Chandra			Rabindranath Tagore			Sarat Chandra Chattopadhyay		
Story name	Genre	P	Story name	Genre	P	Story name	Genre	P
Durgeshnan dini(1865)	Historical Romantic	F-A2	Bou Thakuran ir Haat(1883)	Historical	F-A2	Borodidi (1907)	Social	M-A2
Kapalakundala(1866)	Romantic	F-A2	Chokher Bali (1901)	Social, Romantic	F-A2	Devdas (1914)	Romantic	M-A2
Mrinalini (1869)	Historical Romantic	F-A2	Noukadubi (1906)	Social, Romantic	M-A2	Panditama sai(1914)	Social, Romantic	M-A2
Indira(1873)	Social	F-A2	Gora(1909)	Political, Romantic	M-A2	Pallisamaj (1916)	Social	M-A2
Bishabriksha (1873)	Social	M-A2	Ghare Baire (1916)	Political	M-A2	Chandranath (1916)	Social, Romantic	F-A2
Yugalanguriya(1874)	Romantic	F-A1	Chaturanga (1916)	Social	M-A2	Baikunther Will(1916)	Social	M-A2
Radharani (1876)	Social, Romantic	F-A1	Shesher Kabita (1929)	Romantic	M-A2	Parinita (1916)	Romantic	F-A1
Chandra sekhar(1877)	Social	M-A2	Jogajog(1929)	Social	F-A2	Araksaniya (1916)	Social	F-A3
Rajani(1877)	Romantic	F-A2	Dui bon (1933)	Social	F-A2	Niskriti (1917)	Social	F-A2
Krishnakanter Will(1878)	Social	M-A3	Malancha (1934)	Social	F-A2	Charitrohin (1917)	Social	M-A2
Anandamath (1882)	Political	M-A2	Char Odhhay (1934)	Political	F-A2	Datta(1918)	Social, Romantic	F-A2
Devi Chaudhuri(1884)	Historical Romantic	F-A2				Grihodaho (1920)	Social, Romantic	F-A2
						Dena Paona (1923)	Social, Romantic	F-A2
						Nababidha (1924)	Social, Romantic	M-A2
						Pather Dabi (1926)	Political	M-A2
						Shesprasna (1931)	Social	F-A2

Table 4.9: Political and social events in contemporary time of late nineteenth and early twentieth century [1, 2]

Year	Event
1872	The Brahma Marriage Act passed lifting ban on widow remarriage
1882	Women enter university
1885	Bengal Tenancy Act
1905	First Partition of Bengal; spearheading the 'Quit India' Movement
1906	Bengali Brahma start the Society for the Improvement of Backward Classes, which is the earliest pioneering movement in India dedicated to ameliorating the conditions of Hindu untouchables.
1911	Annulment of Bengal Partition
1912	Imperial capital shifted from Calcutta to Delhi
1914-1918	Fall in jute prices affects the Bengal economy and WW-I effect
1917	Influence of Bolshevism in Bengal intelligentsia
1919	Jallianwala Bagh massacre
1920	Non-cooperation movement against colonial rulers led by Gandhi; and civil disobedience
1930	Salt march movement by Gandhi

# Chapter 5

## Conclusion

In this research, we have performed literary analysis in Bengali literature from two distinct viewpoints employing computational approaches. We have proposed the *Word2vec graph* model, a novel approach to represent a story for literary analysis. We extract various features from the graph structure and use relevant word information to perform unsupervised clustering in author attribution, genre detection, stylochronometry tasks in Bengali literature. We evaluate the performance of *Word2vec graph* with both word unigram TF-IDF and stylometry feature sets. Our model performs the best in genre detection and achieves almost equal or better performance as the unigram feature set in the other task even though it has significantly fewer number selected attributes.

Similarly, we have also investigated social structures in contemporary Bengali literature that witnessed a transitional phase of colonial society. We have modeled the novels of the most influential Bengali writers over a 50 years of time span using character interaction graphs and extract impactful features of each of the novels. Then we have analyzed inter-writer and intra-writer novels in different dimensions that include time-space, age/gender, social structure, protagonists characters, genre, historical true events, etc. We have validated three key research questions on the impact of social structures in Bengali literary fiction. Our study results support that historical events, such as the widow remarriage act and various nationalist movements, influenced contemporary literary fiction. We also observe that, even though female characters have a smaller presence in the character arc, they have been granted considerable importance in the previous literature.

As future work, we aim to improve *Word2vec graph* by incorporating more variations and including most contributing features from other feature sets. Instead of handcrafted features, we will investigate how graph embedding methods can be used to represent the graph and perform other classification tasks. As part of future research in character network analysis, we plan to include more authors, diverse genres, and distinct chronological periods. Apart from the traditional form of literature, folk literature constitutes a considerable portion of Bengali literature. Though it was created by illiterate communities and passed down orally from one



generation to another, it tends to flourish Bengali literature. In the future, we will try to include this less explored literature domain in our research from a computational perspective.

we also strive to include more perspectives such as religion, social divide, etc. in our evaluation in character interaction graphs analysis. We like to improve the character identification and sentiment/emotion association with characters procedure using more advanced techniques. We expect that our study will significantly assist future researchers and writers gain further insights into the connection between contemporary society and the character arc in fiction as well as creating a new dimension in computational research in Bengali literature.

## References

- [1] M. Sinha, *Colonial masculinity: The ‘manly Englishman’ and the ‘effeminate Bengali’ in the late nineteenth century*. Manchester University Press, 2017.
- [2] S. Banerjee, “Marginalization of women’s popular culture in nineteenth century bengal,” *Recasting women: Essays in colonial history*, pp. 127–79, 1989.
- [3] N. Kamal, *The population trajectories of Bangladesh and West Bengal during the twentieth century: A comparative study*. PhD thesis, London School of Economics and Political Science (United Kingdom), 2009.
- [4] T. N. Sadraddinova and K. S. Nasirli, “Literature-mirror of society,” *Научный журнал*, no. 5, pp. 56–59, 2019.
- [5] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, “N-gram-based author profiles for authorship attribution,” in *Proceedings of the conference pacific association for computational linguistics, PACLING*, vol. 3, pp. 255–264, sn, 2003.
- [6] Y. Zhao and J. Zobel, “Searching with style: Authorship attribution in classic literature,” in *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pp. 59–68, Australian Computer Society, Inc., 2007.
- [7] O. Halvani, C. Winter, and A. Pflug, “Authorship verification for different languages, genres and topics,” *Digital Investigation*, vol. 16, pp. S33–S43, 2016.
- [8] M. Koppel and J. Schler, “Authorship verification as a one-class classification problem,” in *Proceedings of the twenty-first international conference on Machine learning*, p. 62, ACM, 2004.
- [9] F. Can and J. M. Patton, “Change of writing style with time,” *Computers and the Humanities*, vol. 38, no. 1, pp. 61–82, 2004.
- [10] C. Klaussner and C. Vogel, “Stylochronometry: Timeline prediction in stylometric analysis,” in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 91–106, Springer, 2015.

- [11] H. Gómez-Adorno, J.-P. Posadas-Duran, G. Ríos-Toledo, G. Sidorov, and G. Sierra, “Stylometry-based approach for detecting writing style changes in literary texts,” *Computación y Sistemas*, vol. 22, no. 1, pp. 47–53, 2018.
- [12] V. G. Ashok, S. Feng, and Y. Choi, “Success with style: Using writing style to predict the success of novels,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1753–1764, 2013.
- [13] S. Maharjan, M. Montes, F. A. González, and T. Solorio, “A genre-aware attention model to improve the likability prediction of books,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3381–3391, 2018.
- [14] H. Alharthi, D. Inkpen, and S. Szpakowicz, “Authorship identification for literary book recommendations,” in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 390–400, 2018.
- [15] J. Egbert, “Style in nineteenth century fiction: A multi-dimensional analysis,” *Scientific Study of Literature*, vol. 2, no. 2, pp. 167–198, 2012.
- [16] M. Wynne, “Stylistics and language corpora,” *Encyclopedia of language and linguistics*. Oxford: Elsevier, 2006.
- [17] G. Oberreuter and J. D. Velásquez, “Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style,” *Expert Systems with Applications*, vol. 40, no. 9, pp. 3756–3763, 2013.
- [18] D. Robson, “Heroes and villains,” *Psychologist*, vol. 29, no. 8, pp. 610–613, 2016.
- [19] D. Dodell-Feder and D. I. Tamir, “Fiction reading has a small positive impact on social cognition: A meta-analysis,” *Journal of Experimental Psychology: General*, vol. 147, no. 11, p. 1713, 2018.
- [20] D. Robson, “Our fiction addiction: Why humans need stories,” *Retrieved on June*, vol. 28, p. 2018, 2018.
- [21] G. F. Simons and C. D. Fennig, “Summary by language size,” *Ethnologue: Languages of the World. 12th ed. Dallas, Texas: SIL International*, 2017.
- [22] T. E. Strahan, “Laurie bauer, the linguistics student’s handbook. edinburgh: Edinburgh university press, 2007. pp. ix+ 387.” *Nordic Journal of Linguistics*, vol. 32, no. 1, pp. 165–174, 2009.
- [23] S. Islam, *Banglapedia: national encyclopedia of Bangladesh*, vol. 3. Asiatic society of Bangladesh, 2003.

- [24] K. Reynolds, *Girls only?: gender and popular children's fiction in Britain, 1880-1910*. Harvester/Wheatsheaf, 1990.
- [25] S. E. Jarrott and B. R. McCann, "Analysis of intergenerational relationships in adolescent fiction using a contact theory framework," *Gerontology & geriatrics education*, vol. 34, no. 3, pp. 292–308, 2013.
- [26] H. White, "The historical text as literary artifact," *Narrative dynamics: Essays on time, plot, closure, and frames*, pp. 191–210, 2002.
- [27] R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y Gómez, and P. Rosso, "Authorship attribution using word sequences," in *Iberoamerican Congress on Pattern Recognition*, pp. 844–853, Springer, 2006.
- [28] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *International conference on artificial intelligence: Methodology, systems, and applications*, pp. 77–86, Springer, 2006.
- [29] F. Leuzzi, S. Ferilli, and F. Rotella, "A relational unsupervised approach to author identification," in *International Workshop on New Frontiers in Mining Complex Patterns*, pp. 214–228, Springer, 2013.
- [30] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
- [31] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 86, 2018.
- [32] J. Worsham and J. Kalita, "Genre identification and the compositional effect of genre in literature," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1963–1973, 2018.
- [33] Y. Sari, M. Stevenson, and A. Vlachos, "Topic or style? exploring the most useful features for authorship attribution," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 343–353, 2018.
- [34] S. Das and P. Mitra, "Author identification in bengali literary works," in *International Conference on Pattern Recognition and Machine Intelligence*, pp. 220–226, Springer, 2011.
- [35] S. Phani, S. Lahiri, and A. Biswas, "Authorship attribution in bengali language," in *Proceedings of the 12th International Conference on Natural Language Processing*, pp. 100–105, 2015.

- [36] H. A. Chowdhury, M. A. H. Imon, and M. S. Islam, "A comparative analysis of word embedding representations in authorship attribution of bengali literature," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2018.
- [37] M. A. Islam, M. M. Kabir, M. S. Islam, and A. Tasnim, "Authorship attribution on bengali literature using stylometric features and neural network," in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT)*, pp. 360–363, IEEE, 2018.
- [38] I. Sen, *Woman and empire: representations in the writings of British India, 1858-1900*, vol. 3. Orient Blackswan, 2002.
- [39] M. Chatterjee, *Women in the Novels of Bankimchandra Chatterjee, Saratchandra Chatterjee and Rabindranath Tagore*. PhD thesis, Saurashtra University, 2009.
- [40] N. Chaudhuri, "Social changes as reflected in bengali literature," *Indian Literature*, vol. 14, no. 2, pp. 41–52, 1971.
- [41] M. A. Quayum, "Hindu–muslim relations in the work of rabindranath tagore and rokeya sakhawat hossain," *South Asia Research*, vol. 35, no. 2, pp. 177–194, 2015.
- [42] R. Das and R. Das, "The nation and the community: Hindus and muslims in the novels of bankim chandra chatterjee," in *Proceedings of the Indian History Congress*, vol. 73, pp. 578–587, JSTOR, 2012.
- [43] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
- [44] D. Bamman, B. O'Connor, and N. A. Smith, "Learning latent personas of film characters," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 352–361, 2013.
- [45] J. Truby, *The anatomy of story: 22 steps to becoming a master storyteller*. Farrar, Straus and Giroux, 2008.
- [46] V. Labatut and X. Bost, "Extraction and analysis of fictional character networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–40, 2019.
- [47] D. Kagan, T. Chesney, and M. Fire, "Using data science to understand the film industry's gender gap," *Palgrave Communications*, vol. 6, no. 1, pp. 1–16, 2020.
- [48] M. M. Lauzen and D. M. Dozier, "Maintaining the double standard: Portrayals of age and gender in popular films," *Sex roles*, vol. 52, no. 7-8, pp. 437–446, 2005.

- [49] B. Kjell, W. A. Woods, and O. Frieder, "Discrimination of authorship using visualization," *Information processing & management*, vol. 30, no. 1, pp. 141–150, 1994.
- [50] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.
- [51] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.
- [52] D. Guthrie, *Unsupervised detection of anomalous text*. PhD thesis, Citeseer, 2008.
- [53] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, pp. 1–29, 2008.
- [54] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Applied intelligence*, vol. 19, no. 1-2, pp. 109–123, 2003.
- [55] T. Urvoy, E. Chauveau, P. Filoche, and T. Lavergne, "Tracking web spam with html style similarities," *ACM Transactions on the Web (TWEB)*, vol. 2, no. 1, pp. 1–28, 2008.
- [56] S. R. Pillay and T. Solorio, "Authorship attribution of web forum posts," in *2010 eCrime Researchers Summit*, pp. 1–7, IEEE, 2010.
- [57] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational linguistics*, vol. 26, no. 4, pp. 471–495, 2000.
- [58] M. F. Amasyalı and B. Diri, "Automatic turkish text categorization in terms of author, genre and gender," in *International Conference on Application of Natural Language to Information Systems*, pp. 221–226, Springer, 2006.
- [59] S. Samothrakis and M. Fasli, "Emotional sentence annotation helps predict fiction genre," *PloS one*, vol. 10, no. 11, p. e0141922, 2015.
- [60] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science*, vol. 5, no. 1, p. 31, 2016.
- [61] S. Kar, G. Aguilar, and T. Solorio, "Multi-view characterization of stories from narratives and reviews using multi-label ranking," *arXiv preprint arXiv:1908.09083*, 2019.
- [62] H. Baayen, H. Van Halteren, and F. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, no. 3, pp. 121–132, 1996.

- [63] J. H. Clark and C. J. Hannon, "A classifier system for author recognition using synonym-based features," in *Mexican International Conference on Artificial Intelligence*, pp. 839–849, Springer, 2007.
- [64] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American society for information science and technology*, vol. 57, no. 3, pp. 378–393, 2006.
- [65] J. J. Jung, E. You, and S.-B. Park, "Emotion-based character clustering for managing story-based contents: a cinemetric analysis," *Multimedia tools and applications*, vol. 65, no. 1, pp. 29–45, 2013.
- [66] S. Gil, L. Kuenzel, and S. Caroline, "Extraction and analysis of character interaction networks from plays and movies," *Retrieved June*, vol. 15, p. 2016, 2011.
- [67] M. C. Ardanuy and C. Sporleder, "Clustering of novels represented as social networks," in *Linguistic Issues in Language Technology, Volume 12, 2015-Literature Lifts up Computational Linguistics*, 2015.
- [68] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Movie analysis based on roles' social network," in *2007 IEEE International Conference on Multimedia and Expo*, pp. 1403–1406, IEEE, 2007.
- [69] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.
- [70] D. Elson, N. Dames, and K. McKeown, "Extracting social networks from literary fiction," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 138–147, 2010.
- [71] M. Elsner, "Character-based kernels for novelistic plot structure," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 634–644, 2012.
- [72] A. Agarwal, S. Balasubramanian, J. Zheng, and S. Dash, "Parsing screenplays for extracting social networks from movies," in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 50–58, 2014.
- [73] S.-B. Park, K.-J. Oh, and G.-S. Jo, "Social network analysis in a movie using character-net," *Multimedia Tools and Applications*, vol. 59, no. 2, pp. 601–627, 2012.
- [74] Q. D. Tran and J. E. Jung, "Cocharnet: Extracting social networks using character co-occurrence in movies.," *J. UCS*, vol. 21, no. 6, pp. 796–815, 2015.

- [75] J. Stiller, D. Nettle, and R. I. Dunbar, “The small world of shakespeare’s plays,” *Human Nature*, vol. 14, no. 4, pp. 397–408, 2003.
- [76] R. Alberich, J. Miro-Julia, and F. Rosselló, “Marvel universe looks almost like a real social network,” *arXiv preprint cond-mat/0202174*, 2002.
- [77] P. M. Gleiser, “How to become a superhero,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 09, p. P09020, 2007.
- [78] H. A. Chowdhury, M. A. H. Imon, and M. S. Islam, “Authorship attribution in bengali literature using fasttext’s hierarchical classifier,” in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, pp. 102–106, IEEE, 2018.
- [79] M. T. Hossain, M. M. Rahman, S. Ismail, and M. S. Islam, “A stylometric analysis on bengali literature for authorship attribution,” in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2017.
- [80] T. Sarkar, “Nationalist iconography: Image of women in 19th century bengali literature,” *Economic and Political Weekly*, pp. 2011–2015, 1987.
- [81] A. Majumder, “Can bengali literature be postcolonial?,” *Comparative Literature Studies*, vol. 53, no. 2, pp. 417–425, 2016.
- [82] S. Muhuri, S. Chakraborty, and S. N. Chakraborty, “Extracting social network and character categorization from bengali literature,” *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 371–381, 2018.
- [83] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [84] L. Stanchev, “Creating a similarity graph from wordnet,” in *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, pp. 1–11, 2014.
- [85] M. Ferial, J. P. Balbin, and F. M. Bautista, “Constructing a word similarity graph from vector based word representation for named entity recognition,” *arXiv preprint arXiv:1807.03012*, 2018.
- [86] S. Argamon and S. Levitan, “Measuring the usefulness of function words for authorship attribution,” in *Proceedings of the 2005 ACH/ALLC Conference*, pp. 4–7, 2005.
- [87] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.



- [88] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [89] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, “graph2vec: Learning distributed representations of graphs,” *arXiv preprint arXiv:1707.05005*, 2017.
- [90] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, “Using of jaccard coefficient for keywords similarity,” in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, pp. 380–384, 2013.
- [91] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [92] S. Lahiri, “Complexity of Word Collocation Networks: A Preliminary Structural Analysis,” in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, (Gothenburg, Sweden), pp. 96–105, Association for Computational Linguistics, April 2014.
- [93] S. Chowdhury and W. Chowdhury, “Performing sentiment analysis in bangla microblog posts,” in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1–6, IEEE, 2014.
- [94] N. I. Tripto and M. E. Ali, “Detecting multilabel sentiment and emotions from bangla youtube comments,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–6, IEEE, 2018.
- [95] D. Das, S. Roy, and S. Bandyopadhyay, “Emotion tracking on blogs-a case study for bengali,” in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 447–456, Springer, 2012.
- [96] A. Das and S. Bandyopadhyay, “Sentiwordnet for bangla,” *Knowledge Sharing Event-4: Task*, vol. 2, pp. 1–8, 2010.
- [97] D. Das and S. Bandyopadhyay, “Developing bengali wordnet affect for analyzing emotion,” in *International Conference on the Computer Processing of Oriental Languages*, pp. 35–40, 2010.
- [98] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, “Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives,” in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 2666–2677, 2016.
- [99] F. Nielsen, “Afinn sentiment lexicon,” 2011.

- [100] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.
- [101] R. Jonker and T. Volgenant, "Improving the hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.
- [102] M. P. Goswami, "Study of pc barua's bengali celluloid version of sarat chandra chattopadhyay's novel devdas," *Mass Communicator: International Journal of Communication Studies*, vol. 12, no. 1, pp. 37–40, 2018.
- [103] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [104] T. M. Cover and J. A. Thomas, "Wiley series in telecommunications and signal processing," in *Elements of information theory*, Wiley-Interscience, 2006.
- [105] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: A review," in *Data Clustering*, pp. 29–60, Chapman and Hall/CRC, 2018.
- [106] R. P. Burns and R. Burns, *Business research methods and statistics using SPSS*. Sage, 2008.
- [107] M. Ghaemi and M.-R. Feizi-Derakhshi, "Feature selection using forest optimization algorithm," *Pattern Recognition*, vol. 60, pp. 121–129, 2016.
- [108] S. Min and J. Park, "Network science and narratives: Basic model and application to victor hugo's les misérables," in *Complex Networks VII*, pp. 257–265, Springer, 2016.
- [109] A. Agarwal, S. Balasubramanian, A. Kotalwar, J. Zheng, and O. Rambow, "Frame semantic tree kernels for social network extraction from text," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 211–219, 2014.
- [110] N. Banik and M. H. H. Rahman, "Gru based named entity recognition system for bangla online newspapers," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pp. 1–6, IEEE, 2018.
- [111] S. A. Chowdhury, F. Alam, and N. Khan, "Towards bangla named entity recognition," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pp. 1–7, IEEE, 2018.
- [112] X. Tang, J. Wang, J. Zhong, and Y. Pan, "Predicting essential proteins based on weighted degree centrality," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 407–418, 2013.

- [113] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [114] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, "Clustering coefficient and community structure of bipartite networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875, 2008.
- [115] K. S. Gupta, *The Philosophy of Rabindranath Tagore*. Routledge, 2016.
- [116] S. Kaviraj, *The Unhappy Consciousness: Bankimchandra Chattopadhyay and the Formation of Discourse in India*. School of Oriental & African Studies University of London, 1995.
- [117] S. Bilston, *The Awkward Age in Women's Popular Fiction, 1850-1900: Girls and the Transition to Womanhood*. OUP Oxford, 2004.
- [118] M. Mamun, N. Yeasmin, and M. Shayeekh-Us-Saleheen, "New historicism and humayun ahmed's jochhona o jononir golpo.," *ASA University Review*, vol. 8, no. 2, 2014.
- [119] R. H. Weisberg, *The failure of the word: The protagonist as lawyer in modern fiction*. Yale University Press, 1984.
- [120] J. Boucher and C. E. Osgood, "The pollyanna hypothesis," *Journal of verbal learning and verbal behavior*, vol. 8, no. 1, pp. 1–8, 1969.
- [121] R. B. Inden and R. W. Nicholas, *Kinship in Bengali culture*. Orient Blackswan, 2005.
- [122] S. Chandra, "Conflicted beliefs and men's consciousness about women: Widow marriage in later nineteenth century indian literature," *Economic and Political Weekly*, pp. WS55–WS62, 1987.

# Appendix A

## Census Report

### A.1 Census Report in Contemporary times

Census reports of British India on a ten year time span are available during the period that we are researching. We get male, female ratio of Bengal province from the reports. We observe age distributions for different range and map them to our proposed A1, A2 or A3 groups. Table A.1 shows brief statistics of this information.

Table A.1: Census Report in Contemporary Time of late Nineteenth and Early Twentieth Century in Bengal [3]. All Values Are Indicated as Percentage (%) Form.

Census year	Male	Female	Age A1	Age A2	Age A3
1871	49.962	50.038	41.457	37.789	20.754
1881	49.97	50.03	45.883	32.078	22.0381
1891	49.751	50.249	29.161	31.394	39.443
1901	50.075	49.925	48.800	30.966	20.233
1911	50.994	49.006	37.11	40.82	22.07
1921	51.23	48.77	34.44	41.35	24.21
1931	50.736	49.264	36.57	42.4	21.03
1941	54.142	45.858	35.86	51.14	13

Generated using Postgraduate Thesis L<sup>A</sup>T<sub>E</sub>X Template, Version 1.03. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on March 19, 2021 at 8:52pm.