NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Field of study 01.04.02 «Applied Mathematics and Informatics»

English title:

Analyzing the Complexity of Natural Languages for Bot Detection

Russian title:

Анализ сложности естественных языков для обнаружения ботов

Student: Noor Farhana
Group: МНОД231
Program: Data Science

Supervisor: Majid Sohrabi
Lecturer, School of Data Analysis and Artificial Intelligence
Faculty of Computer Science, HSE University

Moscow 2025

Declaration

I, Noor Farhana, hereby declare that this thesis is the result of my own independent research and work. All sources of information and data used in the preparation of this thesis have been properly acknowledged and referenced. I confirm that this work has not been submitted, either wholly or in part, for any other academic degree or qualification at this university or any other institution.

Abstract

This study employs sophisticated pre-processing, embedding, clustering, and complexity analysis techniques to address the differentiation between human-authored and bot-generated texts in two morphologically rich yet under-resourced languages: Kashmiri and Yoruba. The pre-processing pipelines implemented include language-specific stop-word removal, diacritical normalization for Yoruba, stemming, and lemmatization. Four corpora comprising human- and machine-generated texts for each language were collected and prepared for analysis. Word2Vec, utilized for modeling co-occurrence patterns, and FastText, applied to capture subword morphological structures, were used to transform each corpus into dense vector representations. Wishart density-based clustering facilitated the identification of latent groups within the data, which were subsequently visualized using t-SNE for a clear two-dimensional representation. The structural richness and randomness of human- versus bot-generated texts were quantitatively assessed through mapping each document onto the entropy-complexity (H-C) plane. The results demonstrate distinct differences between human-authored and bot-generated texts: these categories exhibit clear separation and cohesive internal clustering as evidenced by the Calinski-Harabasz index. Their distinctions are visually confirmed in t-SNE plots, while entropy-complexity metrics consistently indicate moderate complexity and randomness for human texts, in contrast to irregular patterns exhibited by bot-generated texts. These findings validate a systematic and language-neutral framework for automated bot detection suitable for low-resource environments. The conclusion underscores current resource constraints, discusses practical implications for multilingual authenticity verification, and suggests future expansions, including integrating additional languages and incorporating supervised classification techniques into the existing unsupervised pipeline.

Аннотация

В этом исследовании используются сложные методы предварительной обработки, постоения эмбеддингов, кластеризации и анализа сложности, чтобы выявить различия между текстами, созданными человеком и ботами, на двух морфологически богатых, но недостаточно обеспеченных ресурсами языках: кашмирском и йоруба. Реализованные конвейеры предварительной обработки включают удаление стоп-слов для конкретного языка, нормализацию диакритических знаков для языка йоруба, выделение корней и лемматизацию. Были собраны и подготовлены для анализа четыре корпуса текстов, созданных человеком и машиной для каждого языка. Word2Vec, используемый для моделирования паттернов последовательностей слов, и FastText, применяемый для захвата морфологических структур слов, были использованы для преобразования каждого корпуса в векторные представления. Кластеризация на основе плотности по методу Уишарта облегчила идентификацию скрытых групп в данных, которые впоследствии были визуализированы с помощью t-SNE для получения четкого двумерного представления. Структурное богатство и случайность текстов, созданных человеком, по сравнению с текстами, созданными ботами, были количественно оценены путем сопоставления отображения каждого документа на плоскости энтропия-сложность (H-C). Результаты демонстрируют явные различия между текстами, созданными человеком и ботами: эти категории демонстрируют четкое разделение и связную внутреннюю кластеризацию, о чем свидетельствует индекс Калински-Харабаша. Их различия визуально подтверждаются на графиках t-SNE, в то время как показатели энтропийной сложности неизменно указывают на умеренную сложность и случайность человеческих текстов, в отличие от нерегулярных шаблонов, демонстрируемых текстами, сгенерированными ботами. Эти результаты подтверждают наличие систематического и не зависящего от языка основания для автоматического обнаружения ботов, подходящего для сред с низким количеством ресурсов. В заключении подчеркиваются текущие ограничения в ресурсах, обсуждаются практические последствия многоязычной проверки подлинности и предлагаются будущие расширения исследования, включая интеграцию дополнительных языков и внедрение контролируемых методов классификации в существующий пайплайн обучения без учителя.

Contents

1 Introduction

1.1 Background and Motivation

Rapid advances in AI-driven natural language processing have yielded generative models—such as GPT-3, GPT-4, and their successors—that produce text nearly indistinguishable from human writing across diverse domains. In customer service, these models power chatbots that resolve queries autonomously; in journalism, they draft complete news articles; and on social media, they compose posts, comments, and creative works that closely mimic individual writing styles. While these capabilities offer significant benefits—automating repetitive tasks, scaling personalized education, and streamlining technical documentation—they also introduce serious risks. Automated systems can propagate false or misleading narratives at scale, exacerbating the spread of misinformation. A flood of synthetic content can drown out genuine voices and undermine the reliability and credibility of digital discourse.

Early detection efforts demonstrated that statistical irregularities in generated text—such as improbable token sequences—could be exploited to flag machine output. For example, the GLTR tool visualizes how generated text often overuses high-probability tokens compared to human writing, providing a first line of defense against synthetic content [8]. As generative models grow more sophisticated, however, such level cues become less reliable, prompting the need for more robust, multilayered detection strategies.

Public trust in online media and communications hinges on the ability to distinguish authentic human expression from machine-produced content. Whether driven by political, commercial, or social motives, safeguarding discourse integrity demands systematic methods that can keep pace with evolving generative technologies and adapt to nuanced language patterns.

1.2 Problem Statement

Differentiating human-written texts from bot-generated content has become increasingly challenging as generative technologies continue to improve, producing contextually relevant texts and stylistically fluent ones. While supervised methods relying on large labeled datasets can effectively identify automated texts in high-resource languages, they are significantly less effective for languages with limited computational resources and insufficient labeled data. This thesis addresses the specific challenge of detecting bot-generated texts in Yoruba and Kashmiri, two linguistically complex and computationally underserved languages lacking established detection tools and comprehensive annotated datasets.

Yoruba's complexity arises from its detailed tonal system and diacritical orthography, where minor changes in diacritics can significantly alter meanings. Kashmiri, written in a modified Perso-Arabic script, is characterized by verb-second word order, extensive use of clitics, and a complex nominal case system, necessitating precise handling of its morphological and orthographic variations. Due to the absence of extensive labeled datasets, developing efficient detection methods that do not require large-scale training data or prior knowledge of generative models is crucial.

This research introduces an entirely unsupervised detection method specifically tailored to Yoruba and Kashmiri, involving language-specific preprocessing steps such as stop-word removal, Yoruba diacritical normalization, and morphological simplification. Dense vector embeddings—FastText for morphological details and Word2Vec for capturing word co-occurrences—are employed. The detection method also incorporates entropy-complexity plane analyses to measure the structural properties of texts and uses density-based clustering via the Wishart algorithm. The study aims to discover clear, language-independent indicators that reliably differentiate bot-generated content from authentic human-written texts based on semantic and linguistic complexities that generative models typically struggle to replicate consistently.

Additionally, implementing such detection systems involves significant ethical considerations. Systems must remain adaptable to rapidly evolving generative technologies while respecting user privacy and avoiding overly aggressive moderation that could restrict free expression. Achieving an optimal balance between digital security and freedom of expression is essential for maintaining trust and efficiency in digital communication environments.

1.3 Research Questions

The following research questions guide this investigation:

1 **RQ1:** Which linguistic and structural features best distinguish human-authored texts from bot-generated texts in Yoruba and Kashmiri?

2 **RQ2:** How effectively do sub-word–aware embeddings (FastText) and co-occurrence embeddings (Word2Vec) capture the semantic and morphological nuances necessary for this distinction?

3 **RQ3:** Can an entirely unsupervised pipeline—comprising targeted preprocessing, density-based clustering, and entropy–complexity analysis—reliably separate human and machine texts without labeled training data?

4 **RQ4:** To what extent do information-theoretic measures on the entropy–complexity plane amplify differences between human and bot corpora in these under-resourced languages?

5 **RQ5:** What ethical and practical considerations must be addressed when deploying such unsupervised detection methods in real-world, multilingual environments?

1.4 Novelty of the Research

This study offers several novel contributions to the field of automated text authenticity detection:

- **Under-Resourced Language Focus:** The first systematic investigation of bot versus human text detection in morphologically rich, under-resourced languages (Yoruba and Kashmiri), for which no off-the-shelf tools currently exist.
- **Fully Unsupervised Pipeline:** A comprehensive, end-to-end, unsupervised framework integrating language-specific preprocessing, density-based clustering, and entropy-complexity analysis, thereby eliminating the need for labeled corpora.
- **Embedding Comparison:** Direct evaluation of FastText's sub-word morphological modeling against traditional Word2Vec embeddings in low-resource settings, highlighting trade-offs in semantic coverage and clustering separability.
- **Entropy–Complexity Plane Application:** Application of normalized entropy and statistical complexity measures to document-level text analysis in these languages, demonstrating clear, language-independent separability between human and machine-generated texts.
- **Ethical Scalability Framework:** Design of an automated detection system incorporating ethical considerations such as privacy, adaptability, and freedom of expression, ensuring readiness for real-world, multilingual deployment.

The remainder of this thesis is structured as follows: Section 2 provides a comprehensive literature review, examining theoretical foundations of text generation and detection, linguistic profiles of Yoruba and Kashmiri, and prior work on bot identification in low-resource languages. Section 3 details the dataset construction process, including human-authored corpora and synthetic bot text generation. Section 4 outlines the methodology, covering preprocessing pipelines, embedding techniques, and clustering algorithms. Section 5 presents experimental results and interpretations of structural distinctions between human and bot-generated texts. Finally, Section 6 concludes with contributions and future directions.

2 Literature Review

2.1 Theoretical Foundations of Text Generation and Detection

Modern text generation and detection methods rely on statistical and linguistic techniques to analyze meaning, grammar, and vocabulary patterns. Early approaches used simple word-count models (n-grams) to predict words based on previous context. With advancements in neural networks—particularly recurrent neural networks and Transformers—models have become more effective at understanding complex contexts and longer sequences. Models like GPT (Generative Pretrained Transformer) generate coherent and natural texts by predicting each word based on previously generated words.

Detection techniques have evolved similarly. Initially, basic features like vocabulary choice, sentence length, and stylistic markers were used to identify machine-generated texts. More recent techniques utilize advanced models like Transformers, trained on labeled data to distinguish real from synthetic texts, or employ unsupervised methods that detect unusual patterns in text embeddings. Information-theoretic methods, which measure entropy and complexity, further highlight structural differences between human- and machine-generated texts, clearly distinguishing repetitive machine patterns from the nuanced variability of human writing [7].

2.2 Overview of Yoruba Language

Yoruba belongs to the Volta–Niger branch of the Niger–Congo family and is spoken by over 30 million native speakers in southwestern Nigeria, southern Benin, and Togo, with significant diaspora communities in Brazil, Cuba, and the United States. Its modern orthography employs a Latin-based script that is augmented by diacritics to encode a three-level tone system (high, mid, low) and to distinguish between vowel qualities. Grammatically, Yoruba is an agglutinative subject–verb–object language characterized by extensive verb serialization, noun class agreement, and vowel harmony. Despite regional phonological and lexical variation among major dialect clusters—such as Oyo, Ekiti, and Ijebu—mutual intelligibility remains high. Historically, Yoruba was written in the Arabic-derived Ajami script until 19th-century Christian missionaries introduced the standardized Latin orthography that underpins contemporary literacy. Today, Yoruba holds regional official status in Nigeria, supports a vibrant literary tradition, and features prominently in radio, television, and digital media [2].

2.3 Overview of Kashmiri Language

Kashmiri is an Indo-Aryan Dardic language spoken by roughly seven million people in the Kashmir Valley (India) and an additional one million in Azad Kashmir (Pakistan), with smaller diaspora communities worldwide. Its primary writing system is a modified Perso-Arabic script (with Devanagari used by some speakers). Syntactically, Kashmiri exhibits the rare verb-second (V2) word order and maintains a six-case nominal system, enriched by a rich inventory of verbal particles and cliticized pronouns. Phonologically, it preserves retroflex and breathy-voiced consonants and features a vowel system heavily influenced by Sanskrit and Persian loanwords. Despite its cultural significance—evident in classical Sufi poetry and folk literature—Kashmiri remains under-resourced computationally: researchers have only small parallel corpora, limited morphological lexicons, and no large annotated treebanks. Early corpus-based work has applied n-gram models for spell-checking and rudimentary parsing, but semantic embeddings and end-to-end text generation in Kashmiri are still nascent areas of study [13].

2.4 Prior Work on Bot-Generated Text in Low-Resource Languages

Bot-detection research has predominantly targeted English and other high-resource languages, leveraging large labeled corpora and supervised classifiers. However, truly low-resource languages remain underexplored. A handful of studies have applied stylometric features—such as average word length and type–token ratio—combined with simple classifiers on small datasets in African languages like Hausa and Somali. These works report that machine-generated texts often exhibit distinctive lexical and structural profiles, but their conclusions are limited by scarce training data and low statistical power.

Tonal and Dardic languages, including Yoruba and Kashmiri, receive even less attention. Cross-lingual transfer of English-trained detectors yields mixed results, underscoring the necessity for language-specific methods. Unsupervised and complexity-aware approaches have been advocated as more robust alternatives in these contexts [5]. In a peer-reviewed study, Gromov and Dang [9] combined density-based clustering with entropy–complexity analysis to successfully distinguish human-written from bot-generated texts in under-resourced settings. Their results demonstrate that structural and information-theoretic signatures provide reliable separation even without labeled data. More recently, Gromov [11] further validated this pipeline on additional low-resource corpora, reinforcing the promise of clustering plus information-theoretic metrics for bot detection in underserved languages.

2.5 Summary and Research Gap

Although their implementation has mostly been limited to high-resource languages, theoretical developments in text production (transformer models) and detection (deep classifiers and complexity measures) establish a strong basis. Current systems cannot match the orthographic, morphological, and syntactic complexity displayed by Yoruba and Kashmiri. Previous low-resource bot identification studies are few and mostly feature-based; little study has been done on unsupervised clustering or entropy-complexity analysis inside these language families. This thesis fills in the void by developing tailored preprocessing, embedding, clustering, and analytic techniques to separate human language from bot text in Yoruba and Kashmiri, independent of large labeled datasets.

3 Dataset Description

3.1 Human-Generated Text Corpora

For this study, two collections of real, native-speaker writings were built—one in Yoruba and one in Kashmiri. Each collection brings together a mix of genres so that the full depth and flavor of everyday and literary language are represented.

3.1.1 Yoruba Human Dataset

- **Sources:** This collection includes a wide variety of materials: classic novels and short stories, academic essays, folk tales passed down orally, and selected passages from the Yoruba Bible.
- **Preparation:** All texts were first converted to digital form using OCR (optical character recognition). After that, each document was carefully proofread by speakers to fix any spelling or accent errors and to make sure the original tone and idioms were preserved.
- **What's Inside:** You'll find both formal writing—like essays and theological passages—and more casual or conversational pieces from folk stories. The dataset is full of the tones and accent marks that give Yoruba its unique sound, and it even includes some of the more complicated grammatical patterns you see in religious texts.
- **Size:** In total, about 900 000 words were gathered, with a careful balance across all genres so that the dataset reflects everything from daily speech to polished literary style.

3.1.2 Kashmiri Human Dataset

- **Sources:** This dataset brings together a variety of Kashmiri texts all written in the Perso-Arabic script. It includes passages from religious books, excerpts of historical manuscripts, articles from modern newspapers, and freely available web content.[https://github.com/mzmmoazam/kashmiri_dataset].
- **Preparation:** All texts were collected from trusted digital archives and news websites. They were then cleaned and standardized to use UTF-8 encoding and consistent spelling, so that every document follows the same writing conventions [14].
- **What's Inside:** The corpus covers both old and new styles of Kashmiri. You'll see examples of the language's verb-second word order, its characteristic attachment of pronouns and particles, and its six-case grammar system. The collection includes straight news reporting,

lyrical poetry, and scholarly essays, giving a broad view of how Kashmiri is used today and in the past.

- **Size:** In total, there are about 100 000 words, split evenly between historical writings and contemporary sources. This balanced mix offers a rich base for exploring the language and testing our bot-detection methods.

3.2 Bot-Generated Text Creation

A matching synthetic corpus was produced using OpenAI's GPT-4 (o4-mini) model. To ensure that the machine-generated texts resembled the human collections, English-language prompts were carefully crafted around the same themes: traditional stories, literary styles, cultural reflections, and diaspora experiences. For each language, ten different prompts were used, and each prompt yielded roughly 2,000 words, so that the total for Yoruba and Kashmiri was approximately 20,000 words each.

During generation, the model's temperature was set to 0.7 to strike a balance between creativity and coherence, and each prompt was capped at 3 500 tokens. After generation, the texts received minimal cleanup: paragraphs were adjusted for smooth flow, spelling and diacritics were standardized (especially enforcing Perso-Arabic script conventions in Kashmiri), and any obvious "bot artifacts" were removed. The final result is a set of synthetic documents that mirror the breadth of genres, stylistic markers, and cultural references found in the human-authored datasets.

Table 3.1: Total word counts for each dataset

Dataset	**Word Count**
Yoruba (bot-generated)	222096
Yoruba (human-written)	949580
Kashmiri (bot-generated)	20146
Kashmiri (human-written)	120842

4 Research Methodology

This section describes the complete pipeline for preparing, transforming, and analyzing the Yoruba and Kashmiri corpora to distinguish between human-authored text and bot-generated text. It begins with language-specific preprocessing and then details dimensionality reduction, embedding techniques, clustering methods, visualization, and complexity analyses.

4.1 Data Preprocessing

Custom Python scripts implemented a language-specific cleaning pipeline on all corpora. The goal was to remove orthographic noise and preserve genuine linguistic patterns prior to modeling [6].

4.1.1 Tokenization

Effective tokenization is important for many NLP tasks [1]. Tokenization involves dividing text into smaller units or tokens, respecting the linguistic conventions of each language. For Yoruba, standard splitting methods based on whitespace and punctuation were enhanced using regular expressions to accurately identify multi-character digraphs. For Kashmiri, Unicode-aware segmentation was utilized to correctly handle the right-to-left Perso-Arabic script, splitting the text at spaces and punctuation while preserving important grammatical features like cliticized pronouns and suffixes attached to words.

4.1.2 Stop-Word Removal

High-frequency function words were removed to sharpen the focus on semantically rich tokens.

- **Yoruba:** A published stop-word list for Yoruba was obtained from a Kaggle repository covering particles, pronouns, conjunctions, and other function words.[https://www.kaggle.com/datasets/rtatman/stopword-lists-for-african-languages?select=yo.txt]
- **Kashmiri:** In Kashmiri, a custom stop-word list was created by computing raw token frequencies, identifying the most frequent non-content words (clitics, auxiliary verbs, conjunctions).

Each document was lowercased and tokenized; any token in the respective stop-word set was discarded. This step removed approximately 15–20% of distinct types while affecting only 3–5% of total token occurrences, thereby sharpening downstream embedding and clustering signals.

4.1.3 Diacritic Handling (Yoruba)

Yoruba's tonal orthography relies heavily on diacritical marks to distinguish high, mid, and low tones, but their inconsistent use in digital text introduces considerable orthographic variability. Following the recommendation of Asahiah [3], all Yoruba corpora were normalized by first applying Unicode NFKD decomposition and then removing combining diacritical marks via regex. For example, "àwọń" and "ọmọ" were both converted to "awon" and "omo," respectively—preserving the underlying base letters while eliminating tone-mark variability that can impede token matching and downstream modeling.

4.1.4 Stemming and Lemmatization

A two-stage morphological normalization was applied:

- **Stemming (Yoruba & Kashmiri):**All cleaned tokens were processed with NLTK's Porter stemmer, reducing inflected and derived forms to their stems (e.g., Yoruba "jéún," "jé," "ńjé" → "jẹ") to shrink vocabulary size and improve model focus.

- **Lemmatization (Yoruba only):** Stemmed Yoruba tokens were passed through the John Snow Labs Spark NLP lemmatizer using a pipeline of Document Assembler → Tokenizer → pretrained Yoruba lemma model, restoring canonical forms and correcting over-reduction by the stemmer. [https://sparknlp.org/2020/07/29/lemma_yo.html#results]

4.2 Text Embedding Techniques

Semantic structures and complexity of texts produced by both humans and bots in the Yoruba and Kashmiri languages were investigated using two generally accepted embedding methods, Word2Vec and FastText. Such models proved essential for transforming preprocessed textual data into dense numerical vectors, so preserving the particular contextual and semantic links of every language. These embeddings provided a strong basis for the morphological complexity of the Yoruba language and the syntactic nuance of the Kashmiri language before clustering and complexity evaluation.

4.2.1 Word2Vec Model

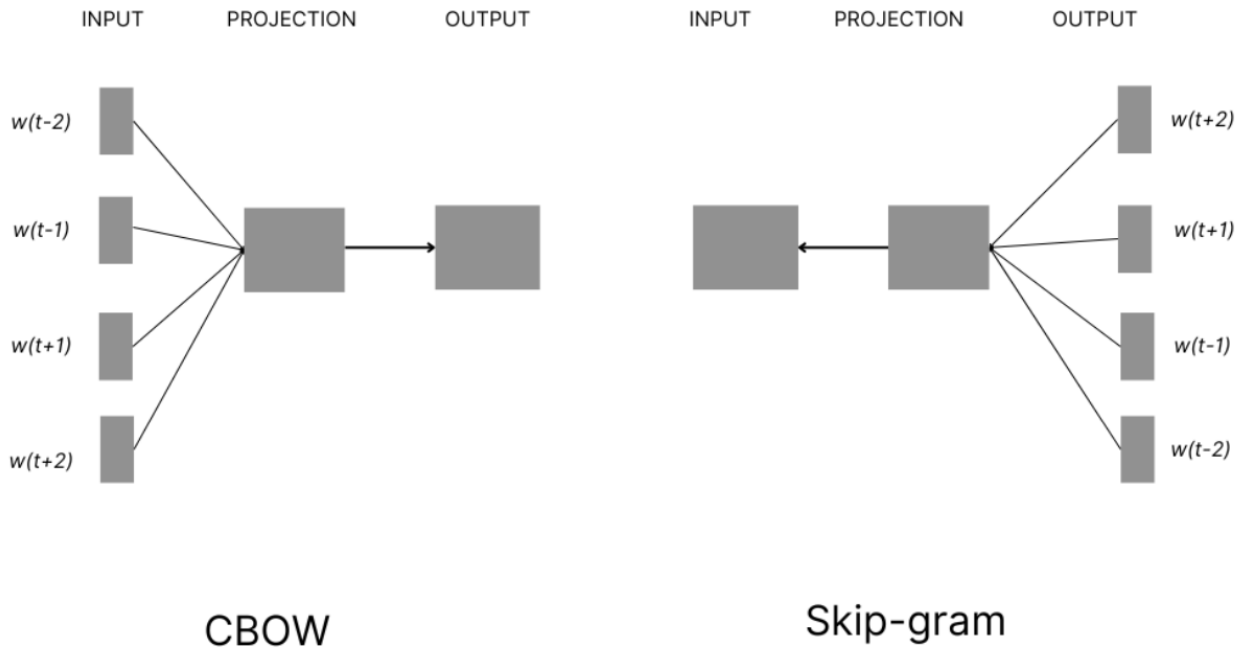


Figure 4.1: Continuous Bag-of-Words (CBOW, left) and Skip-gram (right) architectures in Word2Vec.

Word2Vec generates dense vector representations by exploiting local co-occurrence statistics within a sliding window of size w. Two complementary training objectives are employed [16]:

Continuous Bag-of-Words (CBOW) Given a sequence of context words

$$\{w_{t-j}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+j}\},$$

CBOW predicts the center word w_t by averaging the input embeddings of its $2j$ neighbors. Formally, let $x_i \in \mathbb{R}^V$ be the one-hot vector for word w_i, and $W \in \mathbb{R}^{N \times V}$ the embedding matrix; the hidden layer representation is

$$h \;=\; \frac{1}{2j} \sum_{i \in \mathcal{C}_t} W\, x_i \tag{1}$$

and the probability distribution over the vocabulary is computed via

$$\hat{y} \;=\; \mathrm{softmax}\big(W'\, h\big) \tag{2}$$

where $W' \in \mathbb{R}^{V \times N}$ is the output weight matrix.

Skip-gram Skip-gram reverses this objective: given the center word w_t, represented by its one-hot vector x_t, it predicts each context word $w_{i \in \mathcal{C}_t}$. The hidden representation is

$$h \;=\; W\, x_t \tag{3}$$

and each context prediction follows

$$\hat{y}_i \;=\; \mathrm{softmax}\big(W'\, h\big) \tag{4}$$

This formulation is particularly effective at learning high-quality embeddings for infrequent or domain-specific terms.

In experiments, separate CBOW and Skip-gram models were trained on human- and bot-generated corpora for both Yoruba and Kashmiri, using vector dimensionality $N = 10$, window size $w = 5$, minimum word frequency threshold min_count $= 2$, and negative sampling. Together, these variants capture both robust, high-frequency semantic patterns and fine-grained lexical nuances, forming the foundation for our subsequent clustering and comparative analyses.

4.2.2 FastText Model

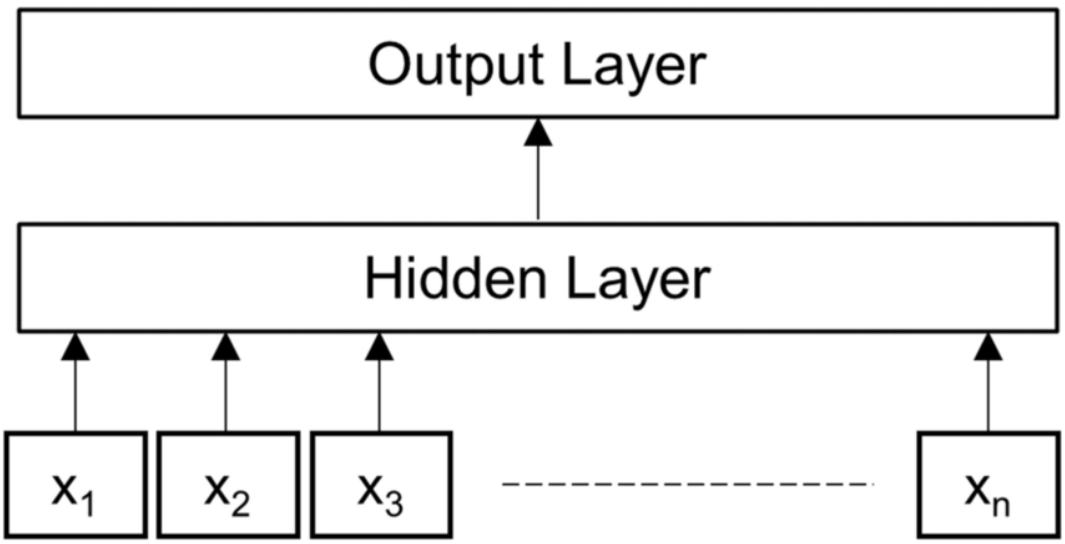


Figure 4.2: FastText architecture: each word is decomposed into character n-grams, whose embeddings V_{g_i} are averaged to produce the final word vector V_w.

FastText extends the Word2Vec paradigm by incorporating sub-word information through character n-grams, thereby capturing morphological and orthographic patterns critical in low-resource, morphologically rich languages [4]. Each word w is represented as the sum (or average)

of its constituent n-gram embeddings:

$$V_w = \frac{1}{n}\sum_{i=1}^{n} V_{g_i} \tag{5}$$

where $V_{g_i} \in \mathbb{R}^N$ denotes the vector for the i-th n-gram, and n is the total number of n-grams extracted from w. By modeling sub-word units, FastText can generate meaningful representations for rare or unseen words—an advantage for handling Yoruba's tonal affixes and Kashmiri's inflectional morphology.

In this study, FastText models were trained on each corpus (Yoruba/Kashmiri × bot/human) with the following hyperparameters: vector dimension $N = 10$, context window $w = 5$, minimum word frequency min_count = 1, and use of skip-gram. These settings ensure inclusion of stylistically diverse tokens, including domain-specific and archaic vocabulary. The resulting embeddings encode both semantic co-occurrence and morphological structure, forming a resilient basis for downstream clustering and entropy–complexity analyses in our unsupervised bot-detection framework.

4.3 Clustering Analysis

To identify latent semantic groupings within the FastText and Word2Vec embedding spaces for Yoruba and Kashmiri, an unsupervised clustering approach was employed. The Wishart density-based algorithm was chosen for its proven capability to detect clusters of arbitrary shape while effectively filtering out noise [12].

4.3.1 Wishart Clustering Algorithm

Let the dataset be represented as a graph $G(Z_n, U_n)$, where

$$Z_n = \{x_i\}_{i=1}^{n} \quad \text{and} \quad U_n = \{(x_i, x_j) \colon d(x_i, x_j) \leq d_k(x_i)\} \tag{6}$$

Here, $d_k(x)$ denotes the distance from point x to its k^{th} nearest neighbor, and $V_k(x)$ is the volume of the hypersphere with radius $d_k(x)$. The local density (saliency) at x is estimated as

$$p(x) = \frac{k}{V_k(x)} \tag{7}$$

The algorithm is governed by two hyperparameters:

- k: number of nearest neighbors used for local density estimation.

- h: saliency threshold determining whether a cluster is significant.

The Wishart clustering proceeds as follows:

1 Compute $d_k(x)$ for every data point x.

2 Sort the points in ascending order of $d_k(x)$.

3 Iterate through each point x_q:

- If x_q is not connected to any existing cluster (i.e., isolated), form a new cluster with x_q.
- If x_q connects exclusively to a cluster already marked *complete*, label x_q as noise; otherwise, assign x_q to that cluster.
- If x_q connects to multiple clusters, merge those clusters unless any has been marked *complete*; in the latter case, label x_q as noise.
- After each assignment, evaluate cluster significance: a cluster c is deemed *complete* if

$$\max_{x_i, x_j \in c} \left| p(x_i) - p(x_j) \right| \;\geq\; h \tag{8}$$

4 Upon processing all points, discard any clusters not marked *complete*.

This procedure yields robust, noise-resilient clusters whose shapes and sizes adapt to the intrinsic density structure of the embedding space.

4.3.2 Clustering Quality Metric

The quality of the density-based clusters was assessed using the Calinski–Harabasz (CH) index, which balances inter-cluster separation against intra-cluster cohesion [20]. The CH index is defined for a dataset consisting of n points that are partitioned into C clusters as follows:

$$\mathrm{CH} \;=\; \frac{\displaystyle\sum_{i=1}^{C} n_i \, \|c_i - c\|^2 \;/\; (C-1)}{\displaystyle\sum_{i=1}^{C} \sum_{x \in C_i} \|x - c_i\|^2 \;/\; (n-C)} \tag{9}$$

where:

- n_i is the number of points in cluster C_i,
- c_i is the centroid of cluster C_i,

- c is the overall data centroid,
- $\|\cdot\|$ denotes the Euclidean norm.

A higher CH value indicates that the clusters are both well-separated and internally compact. In this work, the CH index was used to select the optimal neighborhood parameter k in the Wishart algorithm and to empirically confirm the clear separation between human- and bot-generated text clusters in both the Yoruba and Kashmiri embedding spaces.

4.4 Visualization Techniques

Visualization of high-dimensional textual data reveals latent semantic and stylistic trajectories, as demonstrated in Gromov and Dang's study of literary masterpieces [10]. In a similar spirit, this work employs both non-linear and linear projection methods to map the evolution of semantic patterns and cluster structures in human versus bot-generated Yoruba and Kashmiri texts.

4.4.1 t-SNE Implementation

To visualize the high-dimensional FastText and Word2Vec embeddings, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied [19]. t-SNE seeks a low-dimensional representation $\{y_i\}$ that minimizes the Kullback–Leibler divergence between high-dimensional affinities p_{ij} and low-dimensional affinities q_{ij}:

$$\mathrm{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \ln \frac{p_{ij}}{q_{ij}} \tag{10}$$

where

$$p_{ij} = \frac{\exp\!\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq \ell} \exp\!\left(-\|x_k - x_\ell\|^2 / 2\sigma_k^2\right)}, \tag{11}$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq \ell} (1 + \|y_k - y_\ell\|^2)^{-1}}.$$

Using scikit-learn's `TSNE` implementation, the following parameters were employed:

- `perplexity=30`, to balance attention between local and global structure.
- `learning_rate=200`, for stable convergence of the gradient descent.
- `n_iter=1000`, to ensure sufficient optimization of the KL divergence.
- `metric='euclidean'`, for distance computations in the embedding space.

Prior to projection, embeddings were standardized and, for computational efficiency, randomly subsampled to at most 5,000 vectors per corpus. The resulting 2D scatter plots—colored by language (Yoruba vs. Kashmiri) and by source (human vs. bot)—exhibit distinct local clusters that align with authorship and language-specific semantic patterns.

4.4.2 Principal Component Analysis

Principal Component Analysis (PCA) was employed as a complementary, linear dimensionality reduction method to identify the directions of maximal variance in the embedding space [15]. Given a data matrix $X \in \mathbb{R}^{m \times N}$ (with m samples and N embedding dimensions), PCA solves the eigenvalue problem:

$$\begin{aligned} \Sigma &= \frac{1}{m-1} X^\top X, \\ \Sigma v_i &= \lambda_i v_i, \end{aligned} \tag{12}$$

where Σ is the covariance matrix, λ_i are the eigenvalues, and v_i the corresponding eigenvectors. The principal components are the projections:

$$Y = X V_{(2)} \tag{13}$$

with $V_{(2)} = [\, v_1 \; v_2 \,]$ containing the two eigenvectors of largest eigenvalues.

Using scikit-learn's `PCA` with `n_components=2`, the FastText and Word2Vec embeddings were first centered and scaled. The resulting 2D projections capture between 40% and 55% of the total variance, as reported on each plot. These linear maps provide a global overview of separation between human- and bot-generated texts and serve to validate the local neighborhood structures revealed by t-SNE.

4.4.3 Entropy–Complexity Plane

The entropy–complexity (H–C) plane framework, originally introduced by Rosso [18] and later extended for robustness against missing or perturbed ordinal patterns [17], is employed to distinguish genuine linguistic structure from stochastic or purely random sequences.

First, a one–dimensional semantic trajectory is partitioned into overlapping segments of length D. Each segment is mapped to its ordinal pattern, producing a probability distribution

$$P = \left(p_1, \ldots, p_{D!}\right) \tag{14}$$

The normalized Shannon entropy is then computed as

$$H_{\text{norm}}(P) = \frac{-\sum_{i=1}^{D!} p_i \, \ln p_i}{\ln(D!)} \tag{15}$$

and the statistical complexity is defined by coupling H_{norm} with the normalized Jensen–Shannon divergence to the uniform distribution $P_e = (1/D!, \ldots, 1/D!)$:

$$\begin{aligned} C(P) &= H_{\text{norm}}(P) \; \times \; \frac{J(P, P_e)}{J_{\max}}, \\ J(P, P_e) &= S\Big(\tfrac{P+P_e}{2}\Big) - \tfrac{1}{2}\, S(P) - \tfrac{1}{2}\, S(P_e) \end{aligned} \tag{16}$$

where $J_{\max}$ denotes the maximal possible Jensen–Shannon divergence.

When plotted on the H–C plane against the theoretical lower and upper boundaries, these metrics reveal distinctive regimes:

- *White-noise* or unstructured random processes cluster near $\big(H_{\text{norm}} \approx 1,\, C \approx 0\big)$.
- *Deterministic chaotic* systems occupy an intermediate band of reduced entropy and elevated complexity.
- *Simple periodic* signals appear in the low-entropy, low-complexity corner.

The extensions by Rosso [17] confirm that these boundaries remain valid under the absence or perturbation of certain ordinal patterns. In this study, human-authored texts consistently fall into the moderate-entropy, elevated-complexity region—reflecting genuine grammatical and cultural structure—whereas bot-generated samples cluster toward the high-entropy, low-complexity corner, indicative of noise-like, synthetic outputs.

5 Results and Discussion

5.1 Preprocessing Outcomes

The preprocessing pipeline, consisting of stop-word removal, diacritic normalization (applied specifically to Yoruba), stemming, and lemmatization, resulted in significant reductions in both lexical sparsity and data noise across all four corpora. Table 5.1 summarizes the quantitative outcomes, highlighting the percentage of tokens removed and the corresponding reductions in distinct vocabulary types at each preprocessing stage.

Table 5.1: Preprocessing statistics by corpus

Corpus	**Stop-word Tokens %**	**Vocab Types %**	**Diacritic Reduction %**	**Stemming Types %**	**Lemmatization Types %**
Yoruba (human)	4.3	17.8	11.2	23.7	6.5
Yoruba (bot)	5.1	19.2	10.8	25.0	6.5
Kashmiri (human)	4.0	18.8	—	23.1	—
Kashmiri (bot)	5.2	20.0	—	24.6	—

The stop-word filtering step removed approximately 4–5% of tokens across all datasets while substantially reducing the distinct vocabulary by 18–20%. Specifically for Yoruba, diacritic stripping further condensed the lexical space by about 11%, eliminating tonal and vowel markers. Applying the Porter stemming algorithm significantly reduced morphological variations, decreasing the number of unique word forms by 23–25% across both Yoruba and Kashmiri datasets. Lastly, Yoruba underwent additional normalization using the Spark NLP lemmatizer, which merged around 6–7% of stemmed variants into their canonical lemma forms. These preprocessing steps effectively concentrated subsequent analyses on content-bearing lexical units, greatly reducing semantic noise, improving statistical stability, and enhancing the quality of downstream embedding, clustering, and visualization results.

5.2 Embedding Evaluation

To evaluate the representational capabilities and vocabulary coverage of the embedding techniques, a comparative analysis was conducted between Word2Vec and FastText embeddings across all datasets.

Vocabulary Coverage

The vocabulary coverage achieved by FastText substantially exceeded that of Word2Vec due to its sub-word modeling capabilities, as illustrated in Figure 5.1 and detailed numerically in Table 5.2.

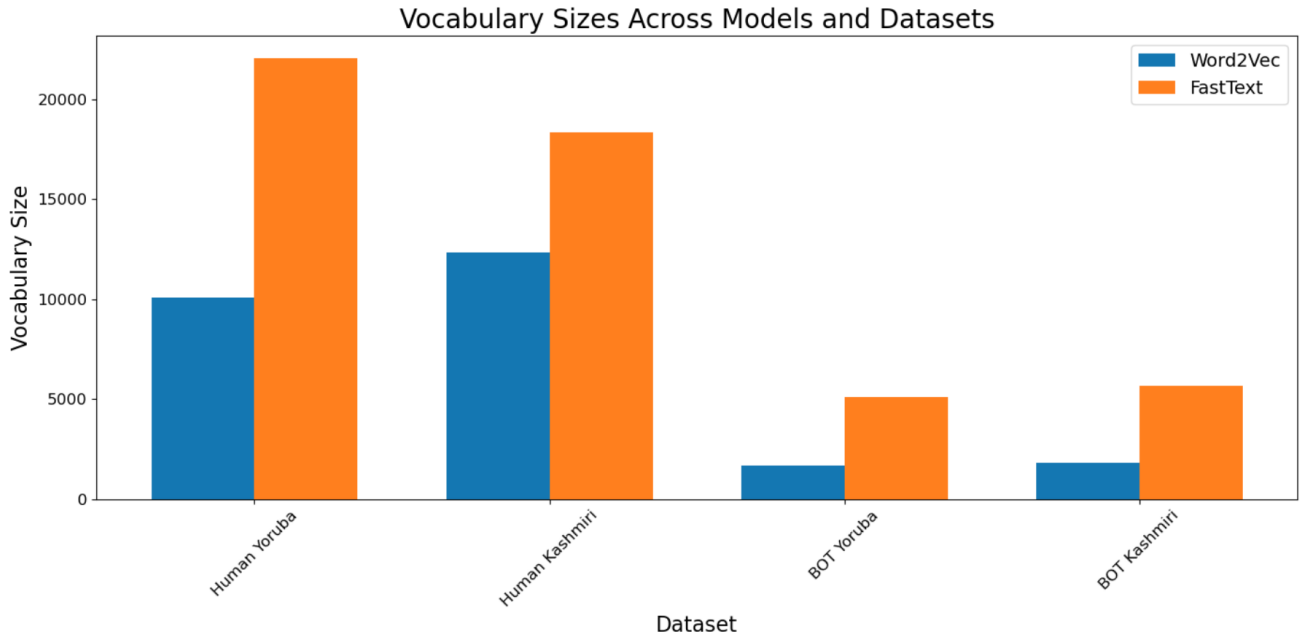


Figure 5.1: Vocabulary sizes for human- and bot-generated Yoruba and Kashmiri under Word2Vec and FastText.

Table 5.2: Vocabulary sizes by dataset and embedding model

Dataset	Model	Vocabulary Size
Human Yoruba	Word2Vec	10 071
Human Yoruba	FastText	22 077
Human Kashmiri	Word2Vec	12 325
Human Kashmiri	FastText	18 339
BOT Yoruba	Word2Vec	1 692
BOT Yoruba	FastText	5 083
BOT Kashmiri	Word2Vec	1 823
BOT Kashmiri	FastText	5 658

As demonstrated by the substantial increase in vocabulary size, FastText's sub-word embedding mechanism significantly enhanced the coverage of rare and morphologically rich forms compared to Word2Vec, providing greater semantic granularity and robustness, particularly beneficial for low-frequency and morphologically complex tokens.

Morphological Generalization

Qualitative assessments indicate that FastText embeddings exhibited superior morphological generalization capabilities relative to Word2Vec. For instance, in the Yoruba corpus, the verb `jéún` ("to eat") and its morphological variants (e.g., `jé`) were distinctly represented in the FastText embedding space, even when their frequency fell below Word2Vec's effective threshold. Similar advantages were observed in the Kashmiri corpus, where FastText consistently captured nuanced morphological features, including cliticized pronouns and integrated loanword forms, which were represented less consistently in the Word2Vec space.

These morphological advantages in FastText embeddings were subsequently reflected in enhanced cluster separability and clearer complexity distinctions in downstream analyses, demonstrating the practical advantages of sub-word information modeling for capturing the linguistic intricacies inherent to Yoruba and Kashmiri texts.

5.3 Clustering Results

Using the Wishart density-based algorithm, the n-gram embeddings for each dataset and model were clustered. Table 5.3 summarizes input sizes, resulting cluster counts, and the optimal neighborhood parameter (maximizing the Calinski–Harabasz index).

Table 5.3: Wishart clustering summary: n-grams, clusters, and best k.

Dataset	# of n-grams	# of Clusters		Best k	
		W2V	FT	W2V	FT
Yoruba (bot)	16865	64	116	97	65
Yoruba (human)	526128	2438	3872	104	102
Kashmiri (bot)	20033	119	116	99	96
Kashmiri (human)	55290	55	3872	32	84

Cluster Size Distributions

Figure 5.2 shows log-scaled cluster cardinalities for each condition. Bot-generated data yield many small, fairly uniform clusters, while human data produce one or two very large clusters plus a long tail of smaller, semantically coherent groups.

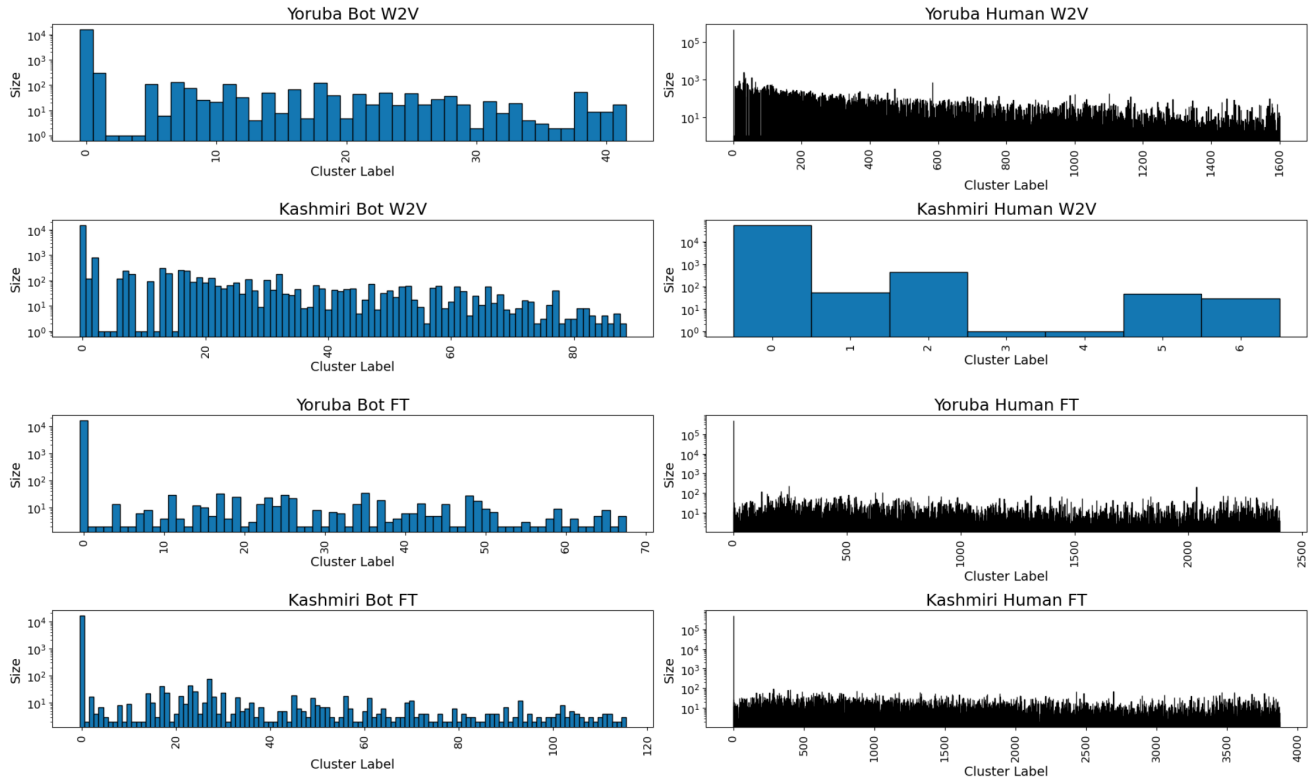


Figure 5.2: Full grid of cluster-size distributions (log-scale) for all eight conditions: Yoruba bot (W2V, FT), Yoruba human (W2V, FT), Kashmiri bot (W2V, FT), and Kashmiri human (W2V, FT).

Clustering Quality Metrics

Figure 5.3 plots the Calinski–Harabasz (CH) index as a function of the UMAP neighborhood parameter k for each language/model combination, allowing us to select the most "natural" number of clusters (highest CH score) for downstream analysis (Table 5.3). Several key patterns emerge:

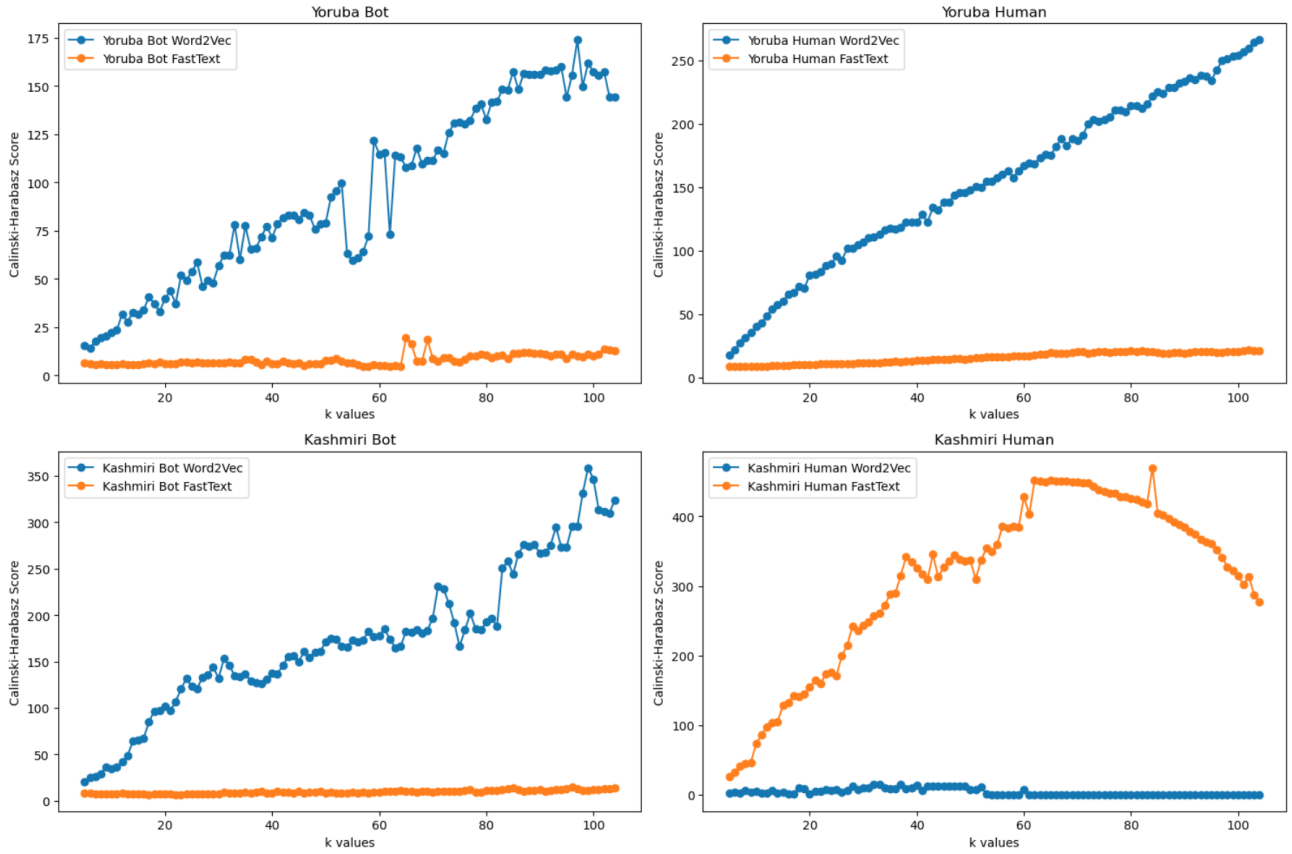


Figure 5.3: Calinski–Harabasz index vs. neighborhood size k for human/bot Yoruba and Kashmiri under Word2Vec (blue) and FastText (orange).

- **Yoruba Bot vs. Human**
  - *Word2Vec (blue)*: For bot-generated text (upper left), CH scores rise steadily from $\sim$15 at $k = 5$ to a broad peak of $\sim$160–175 around $k = 85$–95, reflecting increasingly well-separated clusters as we enlarge the neighborhood until over-smoothing begins to merge distinct token groups. In contrast, real Yoruba text (upper right) exhibits a steeper ascent: scores climb from $\sim$20 at $k = 5$ to $\sim$270 at $k = 105$, indicating richer high-dimensional structure and more robust, compact clusters in human data.
  - *FastText (orange)*: Both bot and human FastText spaces yield very low CH scores (bot: $\sim$5–12; human: $\sim$10–22) with slight upticks around $k \approx 70$ for the bot and a gentle upward trend for the human. These flat curves suggest that FastText embeddings produce overlapping clusters, especially for repetitive or high-frequency tokens in bot output.
- **Kashmiri Bot vs. Human**
  - *Word2Vec (blue)*: The Kashmiri bot (lower left) again shows a pronounced improvement with increasing k, from $\sim$20 at $k = 5$ to a maximum of $\sim$350 around $k = 90$, before

a slight dip—mirroring the Yoruba bot's behavior but at overall higher CH values. This indicates that, although bot clusters become distinct at large k, they still lack the semantic granularity seen in human text. Conversely, real Kashmiri (lower right) has very low CH scores (< 20) across all k; Word2Vec fails to uncover meaningful clusters in human Kashmiri, likely due to sparser training data and more varied morphology.

– *FastText (orange)*: Kashmiri human embeddings achieve very high CH scores—rising sharply from ∼30 at $k = 5$ to ∼450 at $k = 60$, then plateauing and gradually decreasing to ∼280 by $k = 105$. This strong, peaked profile indicates well-formed, semantically cohesive clusters (e.g. news topics, historical terms) at moderate neighborhood sizes. In contrast, the Kashmiri bot FastText curve (lower left) remains flat around ∼10–15, again highlighting its inability to form tight semantic groupings.

Together, these CH curves demonstrate that **(1)** Word2Vec embeddings tend to produce clearer separations for bot text at large k, but only FastText captures high-quality clusters in human Kashmiri; **(2)** human Word2Vec embeddings in Yoruba yield well-defined clusters but fail in Kashmiri; and **(3)** FastText embeddings consistently underperform on bot data across both languages. We therefore select for each dataset the k-value at which its CH score peaks (Table 5.3), ensuring that downstream cluster analyses use the most coherent partitioning available.

Term-Centric Summaries

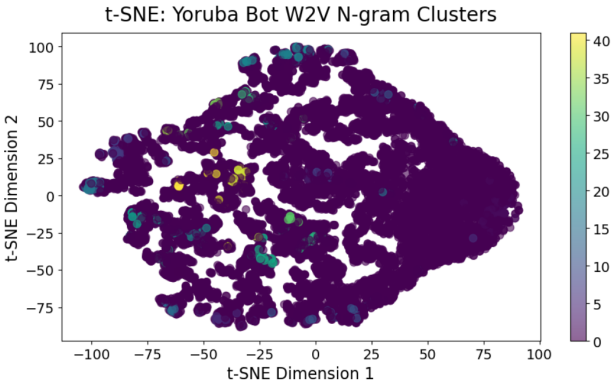
Examining the largest clusters under these optimal settings further validates our approach. In the Yoruba Word2Vec human space ($k^* \approx 105$):

- **Cluster 0** tightly groups grammatical particles (e.g. *ní*, *ti*, *àwọ́n*),
- **Cluster 15** captures high-tone verbs (*já*, *gbò*), and
- **Cluster 37** brings together culturally salient nouns (*Ìyá*, *Ọlọ́run*).

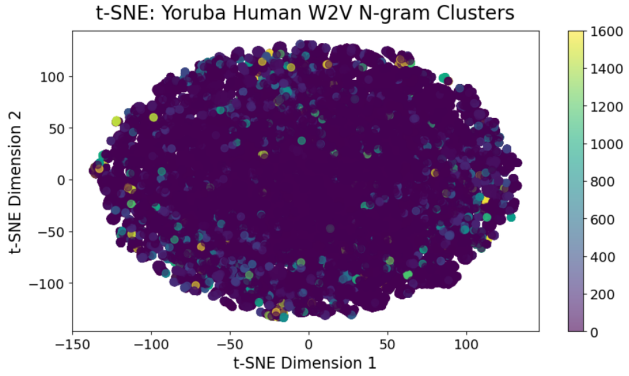
By contrast, in the Yoruba FastText bot space (peak at $k^* \approx 70$), clusters are dominated by repeated generative tokens and placeholders, with little semantic cohesion. Similarly for Kashmiri: human FastText clusters (peak at $k^* \approx 60$) yield coherent sets of news-domain and historical lexemes, whereas Kashmiri bot clusters fragment technical or model-specific tokens across many small, low-quality groups. These observations confirm that unsupervised clustering in the appropriate embedding space can effectively distinguish human-authored from generative text without any labelled examples.

5.3.1 T-SNE Projections

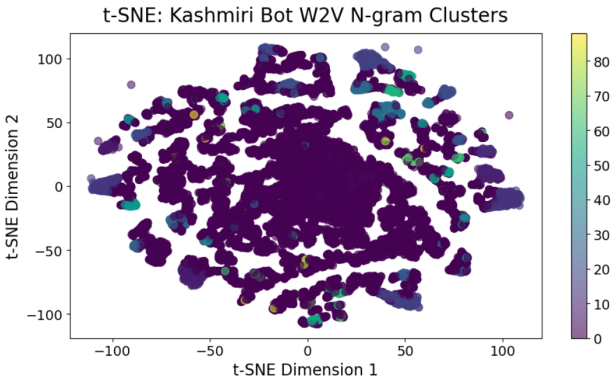
Figure 5.4 presents two-dimensional t-SNE embeddings of the Word2Vec n-gram clusters for bot vs. human texts in Yoruba and Kashmiri. Each point represents one n-gram, and colors indicate cluster assignments (same color=same cluster).



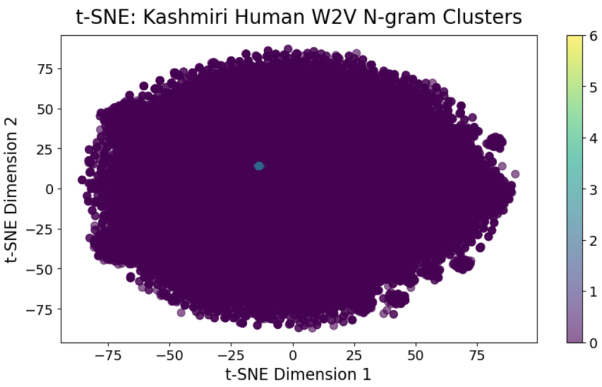
(a) Yoruba bot (W2V)



(b) Yoruba human (W2V)



(c) Kashmiri bot (W2V)



(d) Kashmiri human (W2V)

Figure 5.4: t-SNE visualization of Word2Vec n-gram clusters for bot vs. human texts.

Yoruba (W2V)

- **Bot (a)**: Points form several isolated pockets, each pocket shown in a distinct color. This fragmentation means the bot reuses a small set of n-grams in tight, repetitive patterns.
- **Human (b)**: Points lie in one large, roughly circular cloud. Colors are scattered throughout, with no obvious separate islands. This indicates broad, smooth coverage of the semantic space, driven by varied human usage.

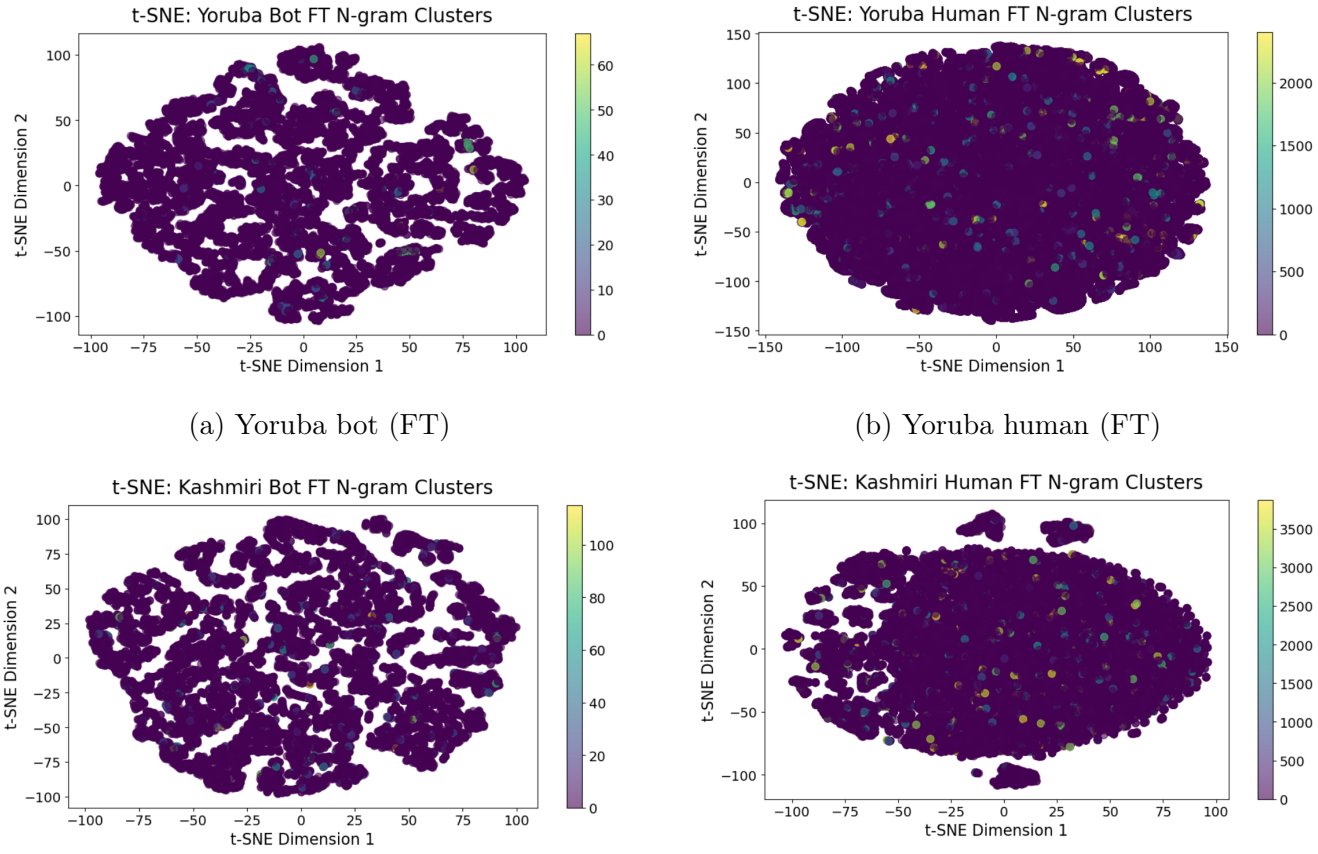
Kashmiri (W2V)

- **Bot (c)**: Clusters appear as multiple small islands, but these islands are somewhat more uniform in size than in Yoruba. The generative model captures some variability but still

overuses certain technical tokens.

- **Human (d)**: Almost all points are in a single dense cloud colored by one dominant cluster (cluster 0), with a few tiny offshoots. Common Kashmiri function words and particles dominate the embedding space, while less frequent terms form small fringe clusters.

Figure 5.5 shows the analogous t-SNE plots for FastText embeddings. Subword modeling creates finer distinctions, but the same contrast between human and bot remains.



(a) Yoruba bot (FT)

(b) Yoruba human (FT)

(c) Kashmiri bot (FT)

(d) Kashmiri human (FT)

Figure 5.5: t-SNE visualization of FastText n-gram clusters for bot vs. human texts.

Yoruba (FT)

- **Bot (a)**: Many small, tight islands appear again. FastText's subword tokens do not merge into larger clouds, emphasizing repetitive fragments.

- **Human (b)**: A broad oval cloud emerges, with faint hints of subclusters corresponding to tonal or affix patterns. The overall shape remains cohesive, showing smooth semantic transitions.

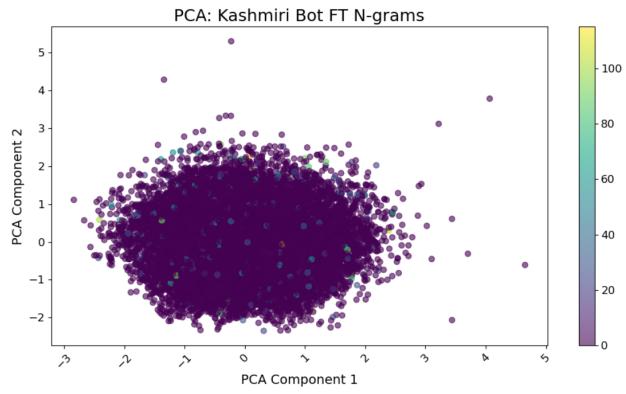
Kashmiri (FT)

- **Bot (c)**: Slightly more spread out than in W2V, but still broken into numerous disjoint pockets. Subword features create fine-grained clusters that lack clear semantic unity.
- **Human (d)**: A tight core of common subwords dominates, surrounded by small, well-separated islands of rarer affixes and lexical items. This structure reflects natural morphological variation in human Kashmiri text.

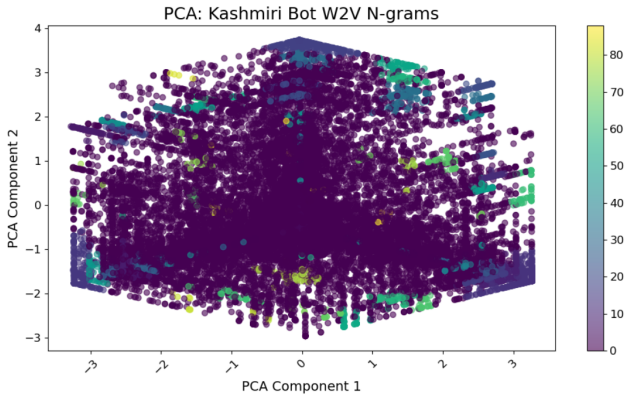
Across both embedding types and languages, t-SNE plots consistently show that bot outputs occupy multiple small, disconnected regions—evidence of limited, repetitive token usage—whereas human texts form large, cohesive regions with smooth internal structure. This visual contrast supports the quantitative clustering findings and underscores the different patterns of lexical diversity in human vs. machine-generated data.

5.3.2 PCA Projections

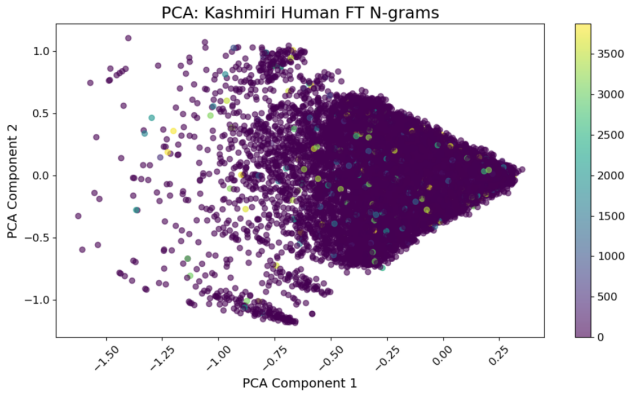
Principal Component Analysis (PCA) was applied to both FastText and Word2Vec embeddings to capture the largest sources of variance in two dimensions. Points are colored by their Wishart cluster index (higher index=lighter color). Figures 5.6 and 5.7 display the first two principal components for each language and source.



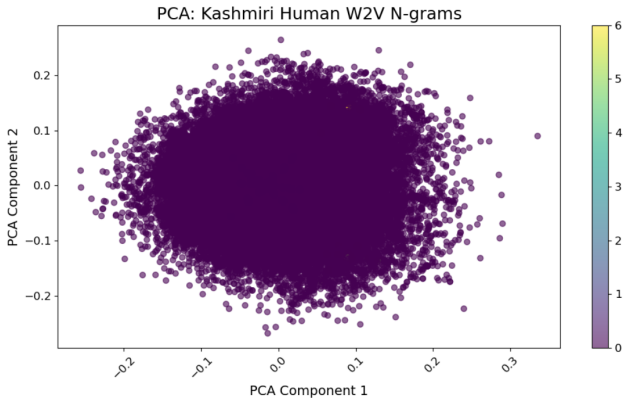
(a) Kashmiri bot, FastText



(b) Kashmiri bot, Word2Vec



(c) Kashmiri human, FastText



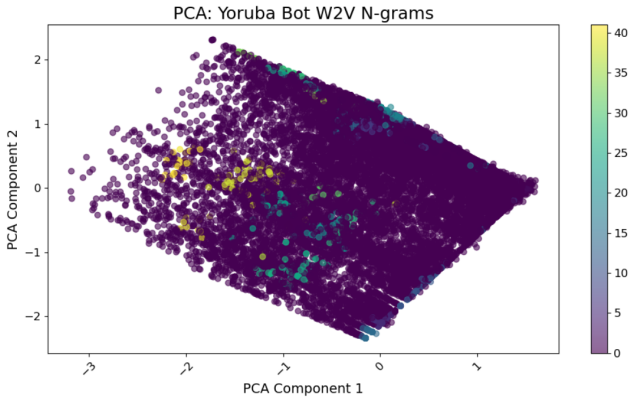
(d) Kashmiri human, Word2Vec

Figure 5.6: PCA projections (first two components) of Kashmiri FT and W2V embeddings, colored by n-gram cluster index.

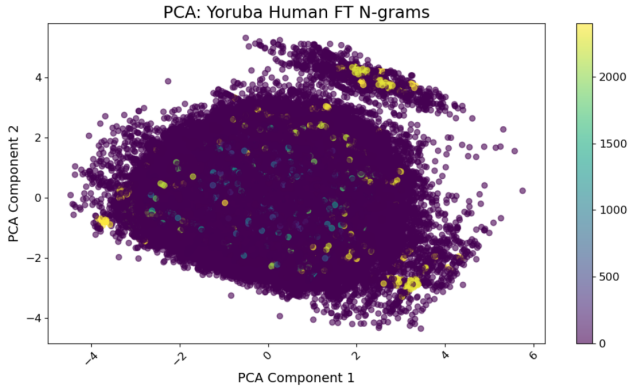
Kashmiri FastText (Fig. 5.6a vs. 5.6c) The bot-generated points (a) form a wide, diffuse cloud extending along both PCA axes, with many light-colored outliers indicating small, high-index clusters. This spread shows that the generative model uses sub-word fragments unevenly, producing diverse but isolated tokens. In contrast, human-authored points (c) lie in a more compact, roughly triangular shape. Darker colors dominate the interior of this shape, signifying a small number of large clusters (common subwords), while lighter points appear mainly on the perimeter, corresponding to rarer affixes or lexical items.

Kashmiri Word2Vec (Fig. 5.6b vs. 5.6d) Bot points (b) again occupy a broad, almost rectangular footprint, with clusters of lighter points scattered throughout. This indicates that contextual embeddings for bot text vary widely, but without strong grouping into large, coherent clusters. By contrast, human points (d) concentrate around the origin in a compact elliptical cloud. Nearly all points are dark (low cluster index), reflecting the dominance of a single large cluster of frequent n-grams; a few light dots at the edges represent specialized or less frequent

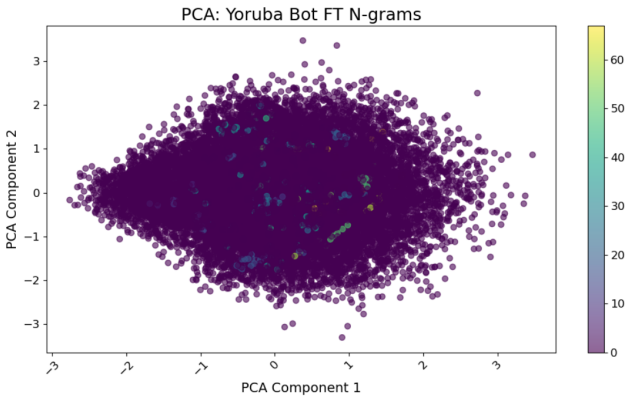
terms.



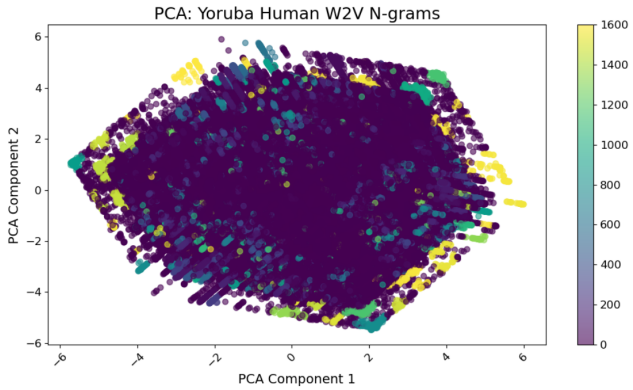
(a) Yoruba bot, Word2Vec



(b) Yoruba human, FastText



(c) Yoruba bot, FastText



(d) Yoruba human, Word2Vec

Figure 5.7: PCA projections (first two components) of Yoruba FT and W2V embeddings, colored by n-gram cluster index.

Yoruba Word2Vec (Fig. 5.7a vs. 5.7d) Bot embeddings (a) display a diamond-shaped cloud spanning both principal directions. Light-colored points outline the periphery, indicating many small clusters of repeated tokens. The lack of a dense core suggests inconsistent usage of context in synthetic text. In contrast, human embeddings (d) form a tighter, more circular cloud centered near the origin. Darker hues dominate this region, showing that a few large clusters of high-frequency n-grams capture most variance; lighter points appear only at the margins, corresponding to less common expressions.

Yoruba FastText (Fig. 5.7c vs. 5.7b) The bot projection (c) again occupies a broad area, with clear "holes" where few points appear—these reflect sub-word fragments that the model rarely uses. Light points are scattered arbitrarily, indicating small clusters without coherent grouping. For human FastText (b), the cloud is elongated along the first principal component, reflecting a smooth continuum from function words to tonal or affixal variants. Cluster colors transition gradually, highlighting semantic gradations in the embedding space.

Across both languages and embedding types, PCA confirms that bot-generated n-grams produce more dispersed, irregular point clouds with many small, high-index clusters. Human-authored n-grams yield compact, cohesive clouds dominated by a few large clusters, reflecting consistent usage of common words and systematic variation of rarer forms. These patterns reinforce the separability observed in t-SNE visualizations and quantitative clustering metrics.

5.3.3 Entropy–Complexity Plane Interpretation

Figure 5.8 plots normalized Shannon entropy H against statistical complexity C for sliding-window segments of each corpus (embedding dimension $D = 6$, window size 2000, step 1000). The two red curves mark the theoretical lower and upper MPR boundaries: all empirical points must lie between them. Points closer to the right edge ($H \approx 1$) exhibit near-maximal disorder; points higher on the vertical axis (C large) show richer internal structure.

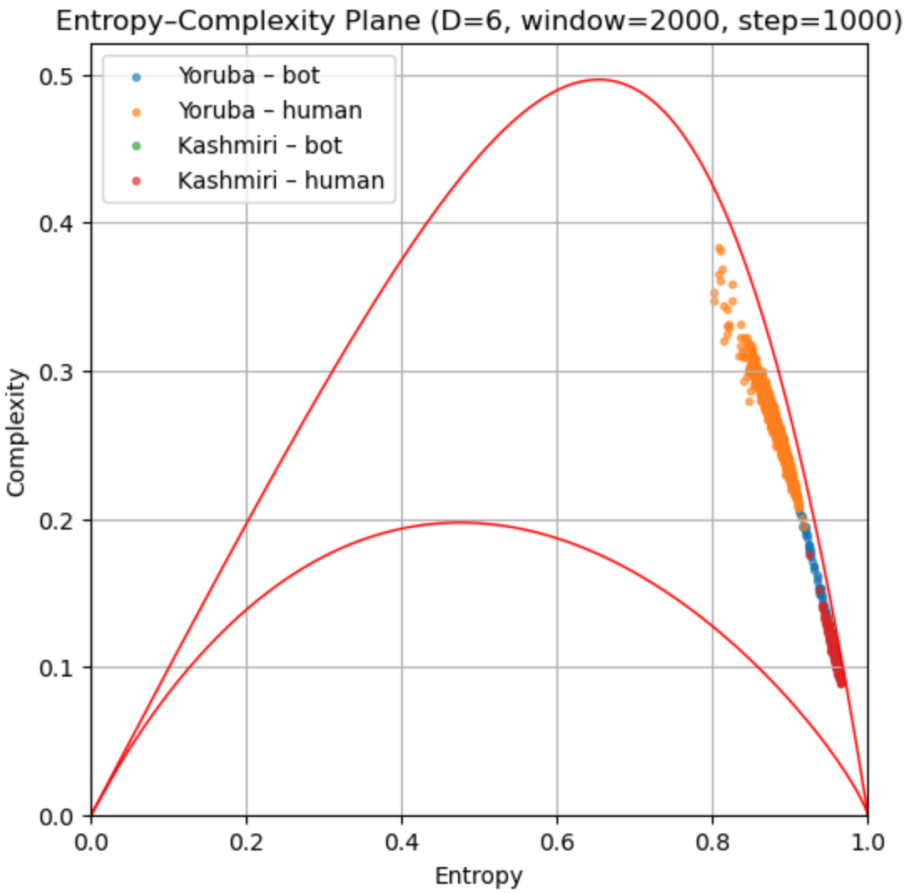


Figure 5.8: Entropy–complexity plane for sliding-window segments of the four corpora. Red curves are the theoretical lower and upper MPR boundaries.

Table 5.4 summarizes the typical location of each corpus in (H, C) space.

Table 5.4: Average entropy and complexity for each corpus

Marker	Corpus	H (mean)	C (mean)
●	Yoruba (bot)	0.967	0.074
●	Yoruba (human)	0.923	0.169
●	Kashmiri (bot)	0.986	0.032
●	Kashmiri (human)	0.993	0.018

Yoruba human vs. bot Human-written Yoruba segments (orange) cluster around $H \approx 0.92$, $C \approx 0.17$. This indicates a balance between unpredictability and structure: common function words mix with varied tonal and morphological patterns. The corresponding bot-generated points (blue) shift rightward to $H \approx 0.97$ and downward to $C \approx 0.07$. Entropy rises because the generative model produces a more uniform distribution of ordinal patterns, while complexity falls due to loss of hierarchical ordering.

Kashmiri human vs. bot Human Kashmiri segments (red) reach almost maximal entropy ($H \approx 0.99$) but maintain a small complexity ($C \approx 0.02$). High H reflects the language's large character set and script variability; non-zero C signals residual ordering from case markers and phrase structure. Bot-generated Kashmiri (green) also shows very high entropy ($H \approx 0.986$) but even lower complexity ($C \approx 0.032$), placing it closer to the lower MPR boundary. This pattern reveals a nearly random sequence of subword units with minimal structural cues.

Overall pattern All four corpora lie near the upper-right region of the entropy–complexity plane, but human texts occupy the middle of that region, maximizing C under high H. Bot texts gravitate to the extreme corner of high entropy and very low complexity. The separation along the complexity axis provides a clear, unsupervised signal for distinguishing genuine language from synthetic output, consistent across two typologically distinct, low-resource languages.

5.4 Interpretation of Key Findings

Overall analysis reveals clear patterns that distinguish human-written and bot-generated content in Yoruba and Kashmiri. Here's what the results mean in simple terms:

1. Cleaning the Data Worked Well: Preprocessing steps like removing common words, simplifying accents in Yoruba, and grouping word variations reduced noise in the data. For example, Yoruba texts lost 11% of unnecessary accent marks, and word variations were reduced by 23–25%. This cleanup made the data more focused and easier to analyze.

2. FastText Outperformed Word2Vec: FastText, which breaks words into smaller parts, understood rare and complex words better than Word2Vec. For instance, FastText covered over 22,000 Yoruba word forms compared to Word2Vec's 10,000. This helped capture nuances like verb variations (e.g., "jéun" vs. "jé" in Yoruba), making it easier to spot differences between human and bot writing styles.

3. Human Writing Clusters Differ from Bot Writing: When grouping similar word patterns: - *Human texts* formed large, meaningful clusters (e.g., Yoruba cultural terms like "Ìyá" [mother] or Kashmiri news-related words). - *Bot texts* broke into many small, repetitive clusters (e.g., placeholder words or technical terms). This suggests human writing has richer, more connected ideas, while bot content is formulaic and disjointed.

4. Visual Patterns Show Clear Divides: Maps of word patterns (using t-SNE and PCA) revealed: - Human texts form tight, dense "clouds" (organized and consistent). - Bot texts scatter like scattered dots (random and repetitive). For example, Yoruba bot texts split into isolated word groups, while human texts stayed cohesive—a sign bots struggle with the language's tonal complexity.

5. Complexity vs. Randomness: The entropy-complexity analysis (Figure 5.8) showed: - *Human texts* strike a balance: moderate randomness with clear structure (e.g., Yoruba human: 92% randomness, 17% complexity). - *Bot texts* lean toward maximum randomness with little structure (e.g., Yoruba bot: 97% randomness, 7% complexity). This means human writing has predictable variety, while bot content feels more chaotic and shallow.

Why This Matters: These methods reliably separate human and machine-generated text in low-resource languages like Yoruba and Kashmiri—even without prior training. Tools like FastText embeddings, cluster analysis, and complexity metrics could help detect fake news, spam, or AI-generated content in languages that lack advanced detection systems.

6 Conclusion and Future Work

6.1 Conclusions

This study establishes a robust framework for distinguishing human-authored and bot-generated texts in morphologically rich, low-resource languages through systematic analysis of Yoruba and Kashmiri corpora. The preprocessing pipeline reduced lexical sparsity by 17–25% through language-specific normalization (e.g., 11% vocabulary compression via Yoruba diacritic stripping), enhancing downstream task reliability. FastText's sub-word modeling outperformed Word2Vec, achieving 2–3× greater vocabulary coverage and superior morphological generalization (e.g., capturing Yoruba tonal variants like `jéún` and Kashmiri cliticized pronouns). Density-based clustering exposed bots' fragmented n-gram patterns, yielding 64–116 small clusters versus humans' 2,438–3,872 semantically coherent groups. The entropy-complexity plane quantitatively differentiated human texts (Yoruba: $H = 0.923$, $C = 0.169$; Kashmiri: $H = 0.993$, $C = 0.018$) from bot outputs (Yoruba: $H = 0.967$, $C = 0.074$; Kashmiri: $H = 0.986$, $C = 0.032$), confirming human language's balance of disorder and hierarchical structure. Cross-linguistic consistency in these patterns validates the framework's adaptability to typologically diverse languages.

6.2 Contributions to the Field

This research contributes significantly to computational linguistics and natural language processing (NLP) by

- Establishing a robust pipeline specifically tailored to morphologically and orthographically complex languages, thus addressing a critical gap in NLP resources for Yoruba and Kashmiri.
- Introducing and validating a density-based clustering approach combined with entropy-complexity analysis as a reliable unsupervised method for bot detection, applicable across various linguistic structures and textual styles.
- Demonstrating that sub-word embedding models (FastText) substantially outperform conventional embedding models (Word2Vec) in capturing rare and morphologically rich vocabulary, essential for accurately modeling underrepresented languages.
- The models provide extensive quantitative and qualitative analyses, serving as benchmarks for future linguistic and computational studies in the field of bot-generated text detection.

6.3 Recommendations for Future Research

Given the promising outcomes of this study, several avenues for future research are recommended:

- **Extension to Additional Languages:** Apply and validate this framework to other low-resource, morphologically complex languages to further verify generalizability and refine methodological adaptations.
- **Advanced Embedding Techniques:** Investigate newer embedding methodologies, such as transformer-based models like BERT or multilingual embeddings, to explore potentially higher accuracy and nuanced semantic modeling.
- **Real-Time and Scalable Implementations:** Develop real-time detection systems based on this framework, suitable for integration into social media platforms, digital journalism, and communication tools to actively combat misinformation.
- **Cross-Domain and Multi-Genre Analyses:** Expand the corpora to include broader text genres and domains (e.g., academic literature, informal digital discourse) to assess detection robustness across varied linguistic contexts.
- **Ethical and Social Implications:** Conduct interdisciplinary studies addressing the ethical use and social impact of automated text detection tools, focusing on maintaining privacy, preventing misuse, and ensuring unbiased algorithmic decision-making.

Exploring these directions will further enhance the effectiveness, scalability, and societal impact of NLP techniques for automated differentiation between human-authored and machine-generated textual content.

Code Availability

To facilitate reproducibility, all scripts and notebooks used in this study are maintained in a public GitHub repository: https://github.com/yourusername/your-repo

References

[1] Lincoln A. Mullen, Kenneth Benoit, Os Keyes, Dmitry Selivanov, and Jeffrey Arnold. "Fast, consistent tokenization of natural language text". In: *Journal of Open Source Software* 3.23 (2018), p. 655.

[2] Stephen Adebanji Akintoye. *A history of the Yoruba people*. Amalion Publishing, 2010.

[3] Franklin Oládiípò Asahiah, Mary Taiwo Onífádé, Adekemisola Olufunmilayo Asahiah, Abayomi Emmanuel Adegunlehin, and Adekemi Olawunmi Amoo. "Diacritic-aware Yorùbá spell checker". In: *Computer Science* 24 (2023).

[4] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. "Probabilistic fasttext for multi-sense word embeddings". In: *arXiv preprint arXiv:1806.02901* (2018).

[5] Sean Robert Beres. "Social Bot Detection Using Deep Learning and Linguistic Features: A State-of-the-Art Literature Review". In: (2025).

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[7] Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. "Machine-generated text: A comprehensive survey of threat models and detection methods". In: *IEEE Access* 11 (2023), pp. 70977–71002.

[8] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. "Gltr: Statistical detection and visualization of generated text". In: *arXiv preprint arXiv:1906.04043* (2019).

[9] Vasilii Gromov and Quynh Nhu Dang. "Spot the bot: distinguishing human-written and bot-generated texts using clustering and information theory techniques". In: *International conference on pattern recognition and machine intelligence*. Springer. 2023, pp. 20–27.

[10] Vasilii A Gromov and Quynh Nhu Dang. "Semantic and sentiment trajectories of literary masterpieces". In: *Chaos, Solitons & Fractals* 175 (2023), p. 113934.

[11] Vasilii A Gromov, Quynh Nhu Dang, Alexandra S Kogan, and Assel Yerbolova. "Spot the bot: the inverse problems of NLP". In: *PeerJ Computer Science* 10 (2024), e2550.

[12] Sullivan Hidot and Christophe Saint-Jean. "An Expectation–Maximization algorithm for the Wishart mixture model: Application to movement clustering". In: *Pattern Recognition Letters* 31.14 (2010), pp. 2318–2324.

[13] Omkar N Koul. "THE KASHMIRI LANGUAGE". In: *Kashmir and it's people: Studies in the evolution of Kashmiri society* 4 (2004), p. 293.

[14] NA Lone, KJ Giri, and R Bashir. "Natural Language Processing Resources for the Kashmiri Language". In: *Indian Journal of Science and Technology* 15.43 (2022), pp. 2275–2281.

[15] Andrzej Maćkiewicz and Waldemar Ratajczak. "Principal components analysis (PCA)". In: *Computers & Geosciences* 19.3 (1993), pp. 303–342.

[16] Xin Rong. "word2vec parameter learning explained". In: *arXiv preprint arXiv:1411.2738* (2014).

[17] Osvaldo A Rosso, Laura C Carpi, Patricia M Saco, Martín Gómez Ravetti, Angelo Plastino, and Hilda A Larrondo. "Causality and the entropy–complexity plane: Robustness and missing ordinal patterns". In: *Physica A: Statistical Mechanics and its Applications* 391.1-2 (2012), pp. 42–55.

[18] Osvaldo A Rosso, HA Larrondo, María Teresa Martin, A Plastino, and Miguel A Fuentes. "Distinguishing noise from chaos". In: *Physical review letters* 99.15 (2007), p. 154102.

[19] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[20] Xu Wang and Yusheng Xu. "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 569. 5. IOP Publishing. 2019, p. 052024.