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, Department 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 or any other institution.

Signed: _____ **Date:** _____

Abstract

This study combines sophisticated pre-processing, embedding, clustering, and complexity analysis techniques to examine the problem of differentiating between human-authored and bot-generated text in two morphologically rich but under-resourced languages: Kashmiri and Yoruba. Our preprocessing pipelines include language-specific stop-word removal, diacritical normalization (Yoruba), stemming, and lemmatization. We also gather four balanced corpora (human and machine texts for each language). We then use Word2Vec (to model co-occurrence) and FastText (to capture subword morphology) to convert each corpus into dense vector representations. We used a method called Wishart density-based clustering to discover hidden groups in the data, and then we used t-SNE to show these groups in a simple two-dimensional view. Lastly, we compare and measure the structural richness and randomness of human versus bot texts by mapping each document to the entropy-complexity (H-C) plane. Our results show a clear difference between texts written by humans and those generated by bots: the two groups are separated and closely grouped together (Calinski-Harabasz index), their shapes are visible in t-SNE plots, and the H–C metrics consistently show that human texts have a moderate level of complexity and randomness, while bot texts have irregular patterns. These results support a methodical, language-neutral framework for automated bot detection in low-resource environments. In our conclusion, we highlight the current resource constraints, talk about the practical implications for multilingual authenticity verification, and outline future directions for adding more languages and adding supervised classification layers on top of our unsupervised pipeline.

Аннотация

Это исследование сочетает в себе сложные методы предварительной обработки, встраивания, кластеризации и анализа сложности, чтобы изучить проблему различия между текстом, написанным человеком и созданным ботом, на двух морфологически богатых, но недостаточно обеспеченных ресурсами языках: кашмирском и йоруба. Наши конвейеры предварительной обработки включают удаление стоп-слов для конкретного языка, нормализацию диакритических знаков (йоруба), выделение основы и лемматизацию. Мы также собираем четыре сбалансированных корпуса (человеческие и машинные тексты для каждого языка). Затем мы используем Word2Vec (для моделирования совпадений) и FastText (для определения морфологии подслова), чтобы преобразовать каждый корпус в плотные векторные представления. Мы использовали метод, называемый кластеризацией на основе плотности по Уишарту, чтобы найти скрытые группы в данных, а затем использовали t-SNE, чтобы отобразить эти группы в простом двумерном представлении. Наконец, мы сравниваем и измеряем структурную насыщенность и случайность текстов, созданных человеком и ботами, сопоставляя каждый документ с уровнем энтропийной сложности (H-C). Наши результаты показывают явную разницу между текстами, написанными людьми, и текстами, созданными ботами: две группы четко разделены и тесно сгруппированы (индекс Калински-Харабаша), их формы видны на графиках t-SNE, а показатели H–C неизменно показывают, что тексты, написанные людьми, имеют умеренный уровень сложности. сложность и случайность, в то время как тексты ботов имеют нерегулярный характер. Эти результаты подтверждают методичную, не зависящую от языка основу для автоматического обнаружения ботов в средах с низким уровнем ресурсов. В нашем заключении мы подчеркиваем текущие ограничения в ресурсах, рассказываем о практических последствиях многоязычной проверки подлинности и намечаем будущие направления для добавления большего количества языков и контролируемых уровней классификации поверх нашего конвейера без контроля.

Contents

1 Introduction

1.1 Background and Motivation

Rapid developments in artificial intelligence-driven natural language processing (NLP) have produced generative models, including GPT-3, GPT-4, and their successors, able to generate text indistinguishable from human-written content across many fields. These models enable chatbots to quickly answer questions in customer support; in journalism, they can independently create whole news stories; and on social media, they produce reasonable posts, comments, and even creative works that mirror personal styles. These features carry great risks even if they bring major advantages, including automating routine tasks, scaling customized educational services, and simplifying technical documentation. Automated systems can quickly spread false narratives, which exacerbates the issue of misinformation. Mass-produced synthetic content could overwhelm real communication, so compromising the dependability and credibility of digital discourse.

Public confidence in digital media and online interactions suffers especially from the consequences since consumers locate it difficult to distinguish real content from synthetic equivalents. Whether politically, economically, or socially driven, protecting online discourse integrity and stopping manipulation calls for aggressive strategies able to regularly separate bot-generated materials from human-authored texts. Establishing these consistent detection techniques becomes a top issue since generative language models get more sophisticated and flexible in reflecting minor language patterns.

1.2 Problem Statement

The difficulty of precisely differentiating between human and bot-generated content gets more difficult as generative bots keep becoming more adept in producing contextually appropriate and stylistically fluid text. While supervised classifiers taught on large labeled datasets can effectively identify subtle markers of automated text generation in high-resource languages, such approaches suffer greatly when used to languages with limited computational resources and inadequate labeled data. This thesis particularly addresses the difficulties of spotting bot-generated texts in two linguistically complex but computationally undersserved languages lacking established detection tools and sufficiently annotated datasets: Yoruba and Kashmiri.

Yoruba presents great difficulties because of its high dimensionality and sensitivity of tokenizing—where small diacritical changes can significantly change meaning—characterized by its sophisticated three-level tonal system and diacritical orthography. Using a modified Perso-Arabic

script, Kashmiri shows a verb-second word order, heavy use of clitics, and a sophisticated six-case nominal system, so requiring exact handling of morphological and orthographic variants. Developing an efficient detection mechanism independent of extensive training data or prior knowledge about particular generative models becomes crucial given the lack of large, labeled datasets for these languages.

This thesis presents an entirely unsupervised detection pipeline specifically for Yoruba and Kashmiri, comprising language-specific preprocessing techniques—such as stop-word filtering, Yoruba diacritical normalisation, and morphological simplification—followed by dense vector embeddings (leveraging FastText for capture sub-word morphological details and Word2Vec for co-occurrence patterns). The pipeline measures information-theoretic properties using entropy–complexity plane analyses and integrates density-based clustering using the Wishart algorithm. This work attempts to find strong, language-agnostic indicators able to differentiate synthetic texts from real human compositions by examining the semantic structures and linguistic complexity generally difficult for generative models to replicate consistently.

Moreover, implementing such a kind of detection system raises important ethical questions. Detection systems have to be flexible enough to avoid fast obsolescence in the face of fast developing generative technologies. Respect of user privacy and avoidance of too harsh moderation that might unintentionally stifle free expression are equally vital. Fostering trust and efficiency in digital environments depends on reaching an ideal balance between digital security and freedom of expression while clearly sharing the detection technique and its constraints.

2 Literature Review

2.1 Theoretical Foundations of Text Generation and Detection

Text generation and detection depend on statistical and linguistic models that fit semantics, syntax, and word frequency. Early methods based on all-occurrence counts estimate the probability of a sequence of words based on generation n-gram language models. More recently, neural sequence models—first recurrent neural networks (RNNs) and then Transformers—had n-grams, rich contextual embeddings, and longer-range dependencies. Transformer-based generators—such as GPT and its variants—predict each token by attending to all previous tokens, so enabling fluid, human-like output in many fields. From basic feature-based classifiers (e.g., bag-of-words, stylometric measures of vocabulary richness and sentence length) to deep neural classifiers using contextual embeddings, detection techniques similarly evolved. Using either supervised labels or

unsupervised anomaly detection on embedding spaces, modern techniques often fine-tune Transformers to discriminate real from synthetic text, where early detectors contrasted surface-level metrics (e.g., lexical diversity, punctuation patterns) between human and machine text. Complementary approaches use information-theoretic metrics—such as Shannon entropy or complexity measures—to quantify randomness and structure, so separating the "over-regularized" patterns of machine output from the subtle irregularities of human writing [5].

2.2 Overview of Yoruba Language

Yoruba is a Volta–Niger language of the Niger–Congo family, spoken by over 30 million native speakers across southwestern Nigeria, southern Benin, and Togo, as well as sizable diaspora communities in Brazil, Cuba, and the United States. It uses a Latin-based orthography enhanced with diacritics to mark its three-level tone system (high, mid, low) and set vowel qualities. Yoruba is an agglutinative, subject-verb-object language with extensive verb serialization, noun class agreement, and vowel harmony. Though they show some phonological and lexical variation, major dialect clusters, including Oyo, Ekiti, and Ijebu, remain generally intelligible. Originally written in the Arabic-derived Ajami script, 19th-century Christian missionaries standardized modern Yoruba handwriting. Yoruba today enjoys regional official status in Nigeria, a rich literary legacy, and widespread use in radio, television, and digital media [1].

2.3 Overview of Kashmiri Language

Kashmiri is an Indo-Aryan Dardic language spoken by approximately 7 million people in the Kashmir Valley of India and by another million in Azad Kashmir, Pakistan, with smaller diaspora communities worldwide. Kashmiri is usually written in a unified Perso-Arabic script (and, to a lesser extent, Devanagari) and features a rare verb-second (V2) word order alongside a six-case nominal system. Features of its syntax are a rich system of verbal particles and a strong system of pronoun clitics. Phonologically, Kashmiri preserves retroflex and breathy-voiced consonants as well as a sophisticated vowel inventory shaped by Sanskrit and Persian loanwords. Although Sufi poetry and folk literature clearly demonstrate the cultural value of Kashmiri, computational resources for the language are scarce: only small parallel corpora, limited morphological lexicons, and no major annotated treebanks are available. Early corpus-based efforts have n-gram models for spell-checking and basic parsing, but semantic embeddings and end-to-end text generation remain developing areas of study [7].

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

Although studies of synthetic text mostly concentrate on English, they also cover many resource-rich languages. Several low-resource studies using stylometric traits and basic classifiers on small datasets have examined African languages, including Hausa and Somali. These studies indicate that human-written and machine-generated texts often have noticeably different average word lengths and vocabulary variety, but not having enough training data usually results in weak overall findings. Bot detection studies for tonal or Dardic languages are few. The mixed results of some studies that train detectors in English and then use them for similar languages show that we need to obtain these languages ready in a special way and use either unsupervised clustering or methods that pay attention to complexity [4].

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

The human-generated corpora assembled for this study encompass authentic, native-speaker texts in Yoruba and Kashmiri languages. These corpora represent a diverse range of genres, thereby capturing the comprehensive linguistic, structural, and stylistic features inherent to each language.

3.1.1 Yoruba Human Dataset

- **Sources:** The corpus includes digitized literary texts, scholarly essays, folk narratives, and selections from the Yoruba Bible.
- **Preparation:** Original documents were digitized via Optical Character Recognition (OCR) and subsequently underwent meticulous manual correction to ensure preservation of orthographic accuracy and idiomatic integrity.
- **Content Profile:** The corpus encompasses formal and semi-formal registers, rich tonal vocabulary, and extensive diacritic-marked lexical forms. Biblical texts specifically enrich the dataset with complex grammatical structures and culturally significant linguistic expressions.
- **Volume:** The final dataset comprises approximately 900,000 words, systematically balanced across multiple genres to accurately reflect both everyday language use and refined literary expression.

3.1.2 Kashmiri Human Dataset

- **Sources:** The corpus integrates religious literature, historical texts, contemporary newspaper articles, and openly accessible digital content predominantly in the Perso-Arabic script.
- **Preparation:** Textual materials were systematically extracted from digital archives and reputable news platforms, followed by normalization procedures to maintain consistent encoding standards and spelling conventions [8].
- **Content Profile:** This dataset captures both classical and modern linguistic registers, highlighting characteristic verb-second word order, complex cliticization, and diverse case inflections. The selection includes varied journalistic language, poetic compositions, and scholarly prose.
- **Volume:** This dataset contains approximately 100,000 words, evenly distributed between historical texts and contemporary sources.

3.2 Bot-Generated Text Creation

- **Model and API:** The synthetic corpora were generated using the GPT-4 (o4-mini) model, accessed via the OpenAI API.

- **Prompt Design:** Prompts written in English explicitly targeted thematic domains that mirrored the human-authored texts, including oral traditions, literary forms, cultural narratives, and diaspora experiences.

- **Generation Parameters:**

  - A total of ten distinct thematic prompts were developed per language.
  - Each generated text comprises approximately 2,000 words, resulting in approximately 20,000 words per language.
  - The temperature parameter was set to 0.7, with a token generation limit of 3,500 per prompt.

- **Post-Processing:** Generated texts underwent minimal processing to ensure orthographic consistency, specifically adhering to the Perso-Arabic script for Kashmiri. Additionally, post-processing maintained paragraph coherence and eliminated noticeable model-generated artifacts.

- **Outcome:** The resultant synthetic corpora were culturally and stylistically aligned to parallel the diverse linguistic features and genres represented within the corresponding human-generated 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

We standardized and cleaned the raw texts before any modeling to make sure that later studies show real language patterns instead of orthographic noise or formatting errors. Every stage was carried out in custom Python scripts written for every language.

4.1.1 Tokenization

We first split the texts into tokens using language-aware rules. For Yoruba, we applied whitespace and punctuation splitting augmented with regular expressions to to serve multi-character digraphs (e.g., "gb," "sh"). For Kashmiri, we used Unicode-aware segmentation to correctly handle the Perso-Arabic script, splitting on spaces and punctuation while retaining attached clitics.

4.1.2 Stop-Word Removal

To focus on content-bearing words, we used high-frequency function tokens:

- **Yoruba:** We employed asive stop-word list sourced from an established Yoruba language repository, covering particles, pronouns, conjunctions, and other common function words. https://www.kaggle.com/datasets/rtatman/stopword-lists-for-african-languages?select=yo.txt.
- **Kashmiri:** We created our own list by computing raw token frequencies, identifying the mostent non-content words (clitics, auxiliary verbs, conjunctions), and manually filtering out any remaining substantive items.

Each document was lowercased and split into tokens; any token appearing in its language's stop-word set was discarded. This process eliminated roughly 15–20% of the distinct vocabulary—while only affecting 3–5% of total word occurrences—thereby sharpening the signal for downstream embedding and clustering.

4.1.3 Diacritic Handling (Yoruba)

Yoruba's tonal orthography relies on diacritics, which introduce variability that hinders token matching. We applied Unicode NFKD decomposition followed by regex-based removal of combining diacritical marks. For example, àwn became awon, and m became omo, preserving the base character forms while eliminating inconsistent tone annotations [2].

4.1.4 Lemmatization and Stemming

Our approach of two-stage morphological normalizing was:

- **Stemming (Yoruba Kashmir:)** All cleaned texts underwent NLTK's Porter stemmer. This shortened inflected and derived forms to their stems (e.g., Yoruba "jun, "j, "ńj," → "j"; Kashmiri "likhmut, "likhnas, "likhan," → "likh"; greatly shrinking the vocabulary and improving model focus.

- **Lemmatization (Yoruba only**: We routed the stemmed Yoruba texts through the John Snow Labs Spark NLP lemmatizer to maintain linguistic precision in Yoruba. Each stemmed token was mapped to its canonical dictionary form using a Spark pipeline (Document Assembler → Tokenizer → Presrained Yoruba lemma model), so correcting over-reduction by the stemmer and so restoring differences required for accurate semantic representation. 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 gave a strong basis for the morphological complexity of the Yoruba language and the syntactic nuance of the Kashmiri language before clusterings and complexity evaluation.

4.2.1 Word2Vec Model

Word2Vec uses word co-occurrence inside a defined window to convert every corpus into dense vectors. Two variants trained independently on human and bot texts for both Yoruba and Kashmiri (vector size = 10, window = 5, min_count = 2):

- **Continuous Bag-of-Words (CBOW)** predicts the target word by averaging the vectors of surrounding context words. It works well for simulating often occurring words.

- **Skip-gram** It looks at the target word as a guide for surrounding context words. Learning representations for rare or domain-specific words is especially where this approach is most successful.

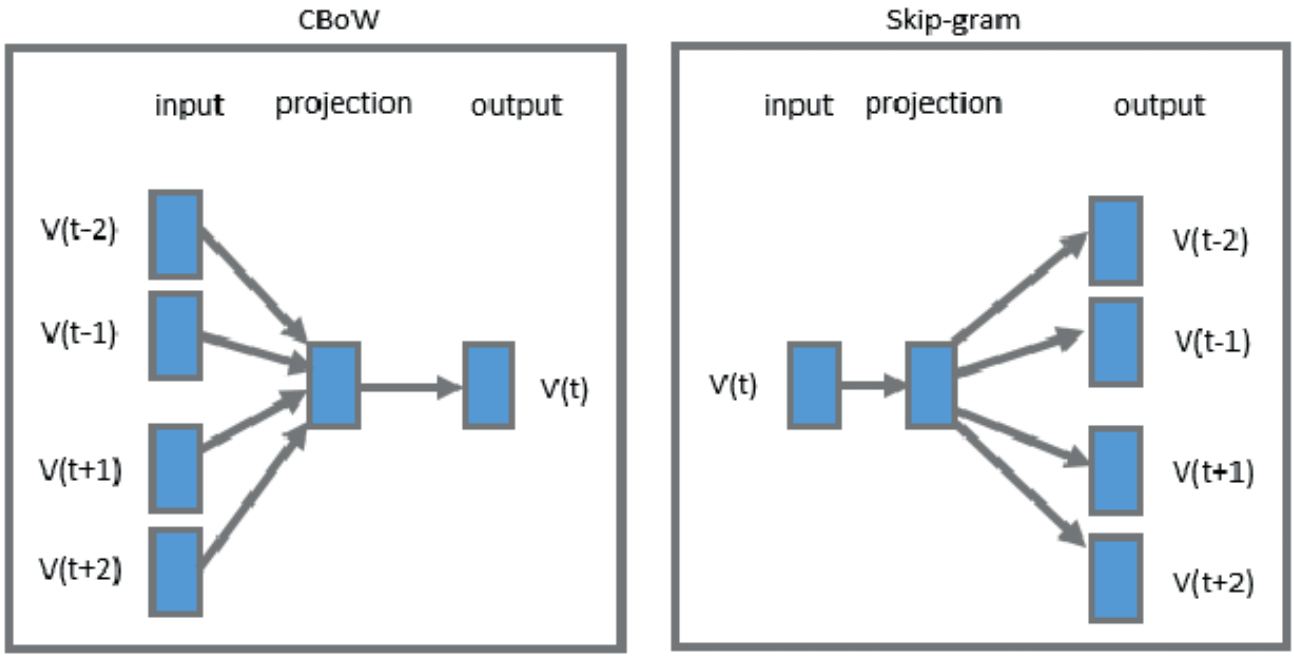


Figure 4.1: CBOW (left) predicts a target word from its context; Skip-gram (right) predicts context words from a target word.

These embeddings taken together capture both common patterns and rare lexical nuances, so offering a strong basis for later clustering and comparison of human against machine text [10].

4.2.2 FastText Model

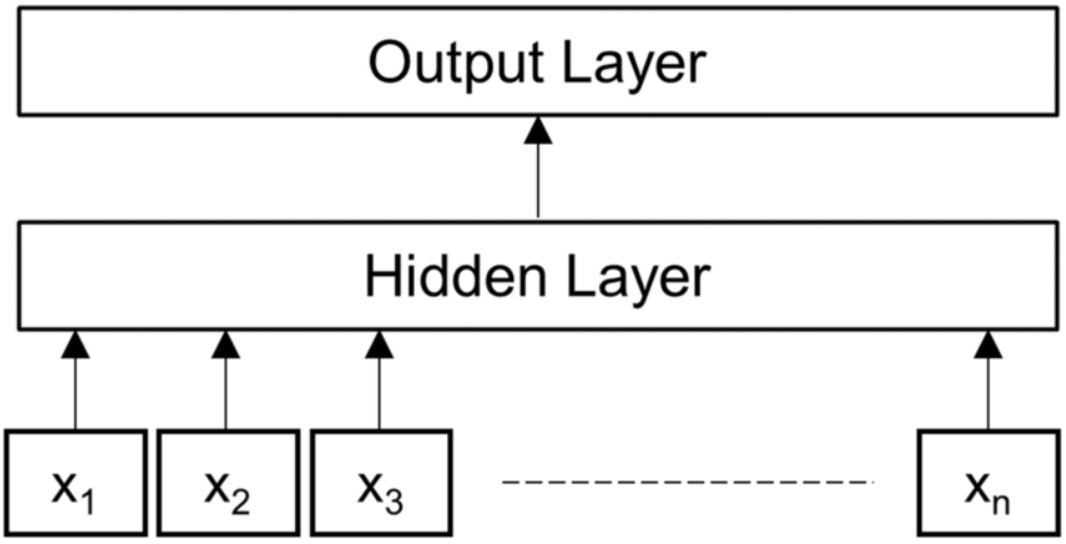


Figure 4.2: FastText architecture: words are decomposed into character n-grams, each embedded and averaged to form the final word vector.

FastText enhances Word2Vec by representing each word as the average of its character

n-gram embeddings:

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

where V_{g_i} is the embedding of the i-th n-gram and n the total count of n-grams in word w. This sub-word modeling is essential for Yoruba's tonal affixes and Kashmiri's complex inflections, enabling reliable encoding of rare or unseen forms.

Models were trained with vector size=10, context window=5, and minimum word frequency=1—parameters chosen to include stylistically distinct and domain-specific vocabulary from literary and scriptural texts. The resulting embeddings capture both semantic and morphological nuances, providing a robust foundation for subsequent clustering and entropy-complexity analyses in the bot-detection framework [3].

4.3 Clustering Analysis

Clustering was applied to the embedding spaces (FastText and Word2Vec) of the Yoruba and Kashmiri corpora to uncover natural groupings without supervision. Wishart's density-based algorithm was selected for its ability to filter noise and form clusters of arbitrary shape [6].

4.3.1 Wishart Clustering Algorithm

Model the dataset as a graph $G(Z_n, U_n)$, where:

$$Z_n = \{x_i\}_{i=1}^{n}, \quad U_n = \{(x_i, x_j) : d(x_i, x_j) \leq d_k(x_i)\},$$

with $d_k(x)$ the distance to the k-th nearest neighbor and local density

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

where $V_k(x)$ is the hypersphere volume of radius $d_k(x)$. Two hyperparameters control the process:

- k: number of neighbors for density estimation
- h: saliency threshold for cluster significance

Algorithm steps:

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

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

3 For each point x_q:

- If isolated, start a new cluster.
- If connected only to a "complete" cluster, label it noise; otherwise assign it there.
- If connected to multiple clusters, merge them if none are marked complete; otherwise label x_q noise.
- Mark a cluster c as complete if

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

4 Discard any cluster not marked complete after processing all points.

4.3.2 Clustering Quality Metric

Cluster validity was measured using the Calinski–Harabasz (CH) index [13]:

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

where n is the total number of points, C the number of clusters, n_i and c_i the size and centroid of cluster i, and c the overall centroid. A higher CH score indicates more distinct, cohesive clusters. In this study, CH guided the choice of k and confirmed clear separation between human- and bot-generated texts in both Yoruba and Kashmiri corpora.

4.4 Visualization Techniques

4.4.1 t-SNE Implementation

t-SNE (t-distributed Stochastic Neighbor Embedding) was used to project high-dimensional embedding vectors into two dimensions while preserving local neighborhood structure [12]. The following settings were applied using scikit-learn's `TSNE` class:

- `perplexity=30` to balance attention between local and global aspects of the data.
- `learning_rate=200` for stable convergence.
- `n_iter=1000` to allow sufficient optimization of the KL-divergence objective.

- `metric='euclidean'` to measure distances in the embedding space.

Before fitting, embeddings from FastText and Word2Vec were standardized and, if necessary, randomly subsampled to 5,000 vectors per corpus for performance. The resulting 2D scatter plots, colored by language and by human vs. bot source, reveal clear local clusters that correspond to authorship and language-specific semantic groupings.

4.4.2 Principal Component Analysis

Principal Component Analysis (PCA) was applied as a complementary linear projection to capture the directions of maximum variance [9]. Using scikit-learn's `PCA` with `n_components=2`, the top two principal axes were extracted from the combined FastText embeddings. Prior to PCA, each embedding dimension was centered and scaled. The percentage of variance explained by these two components (typically 40–55% across corpora) was reported on each plot. PCA maps provide a global view of separation between human and machine text and help validate t-SNE's local cluster structures.

4.4.3 Entropy-Complexity Plane

The entropy–complexity (H–C) plane quantifies each document's randomness and structural richness [11]. Steps:

1 **Symbolic Mapping**: Each cleaned document was converted into a sequence of overlapping character tri-grams.

2 **Shannon Entropy** (H):

$$H = -\sum_i p_i \log_2 p_i,$$

where p_i is the relative frequency of the i-th tri-gram.

3 **Statistical Complexity** (C): López-Ruiz–Mancini–Calbet (LMC) complexity,

$$C = H \times D, \quad D = \sum_i (p_i - \tfrac{1}{N})^2,$$

with N the number of unique tri-grams.

4 **H–C Plotting**: Each document is represented as a point (H, C) in the plane. Human texts occupy a distinct moderate-entropy, high-complexity region, while bot texts cluster either at high-entropy/low-complexity or low-entropy/high-complexity, reflecting over-regularization or repetitive patterns.

5 Results and Discussion

5.1 Preprocessing Outcomes

The cleaning pipeline—stop-word removal, diacritic normalization (Yoruba only), stemming, and lemmatization—yielded substantial reductions in both noise and vocabulary sparsity across all four corpora. Table 5.1 summarizes the percentage of tokens removed and the reduction in distinct word types at each major step.

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	—

After stop-word filtering, roughly 4–5 % of the total tokens were discarded, while the distinct vocabulary shrank by 18–20 %. Diacritic stripping in Yoruba collapsed about 11 % types by removing tone and vowel markers. Applying the Porter stemmer reduced Yoruba and Kashmiri type counts by approximately 23–25 %. Finally, the Spark NLP lemmatizer for Yoruba merged an additional 6–7 % surface forms back into their canonical lemmas, balancing over-reduction and preserving necessary distinctions. These steps concentrated the models on content-bearing vocabulary, improved statistical stability, and substantially lowered noise for all downstream embedding, clustering, and visualization tasks.

5.2 Embedding Evaluation

To assess the representational capacity of our embeddings, we first compare the vocabulary sizes learned by Word2Vec and FastText on each corpus.

Vocabulary Coverage

Table 5.1 shows that FastText's sub-word modeling substantially increases coverage in both Yoruba and Kashmiri:

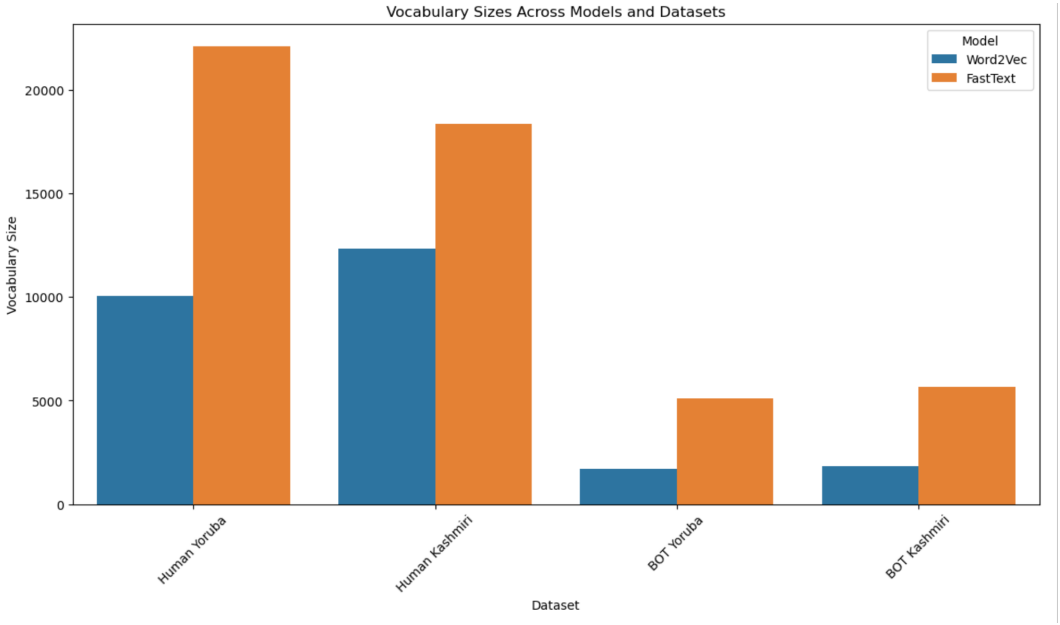


Figure 5.1: Vocabulary sizes for human- and bot-generated Yoruba and Kashmiri under Word2Vec and FastText.

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

Morphological Generalization

FastText shows quite good handling of rare and morphologically complicated forms qualitatively. FastText assigns different vectors to the Yoruba verb `jun` ("to eat") and its inflected variants `j`, even if their frequency threshold falls below Word2Vec's. Likewise, the FastText space shows loanword forms and Kashmiri cliticized pronouns (e.g., `chu`, followed by verb suffixes) more regularly. The enhanced separability shown in our downstream clustering and complexity studies is derived from this robustness in sub-word modeling.

5.3 Clustering Results

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

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

Dataset	#n-grams	#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.

Clustering Quality Metrics

We tracked the Calinski–Harabasz score as a function of k for each dataset and model (Figure 5.3). The score peaks determine our final k-values (Table 5.3).

Term-Centric Summaries

Inspection of the largest human clusters reveals coherent groupings: in the Yoruba Word2Vec space, cluster 0 contains particles (*ní*, *ti*, *awn*), cluster 15 groups tonal verbs (*j*, *gb*), and cluster

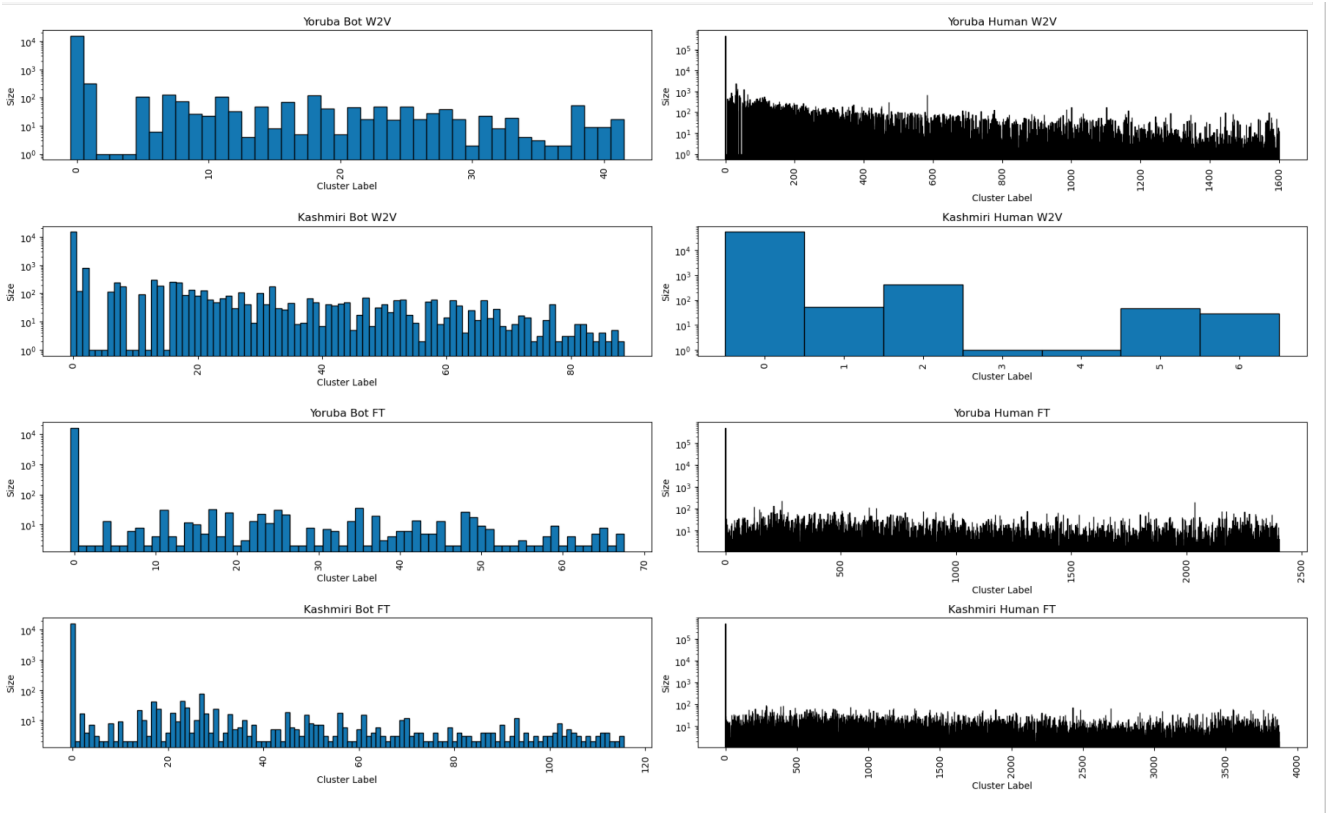


Figure 5.2: Full grid of cluster-size distributions (log-scale) for all eight conditions: Yoruba bot (W2V, FT), Yoruba human (W2V, FT), Kashmiri bot (W2V, FT), and Kashmiri human (W2V, FT).

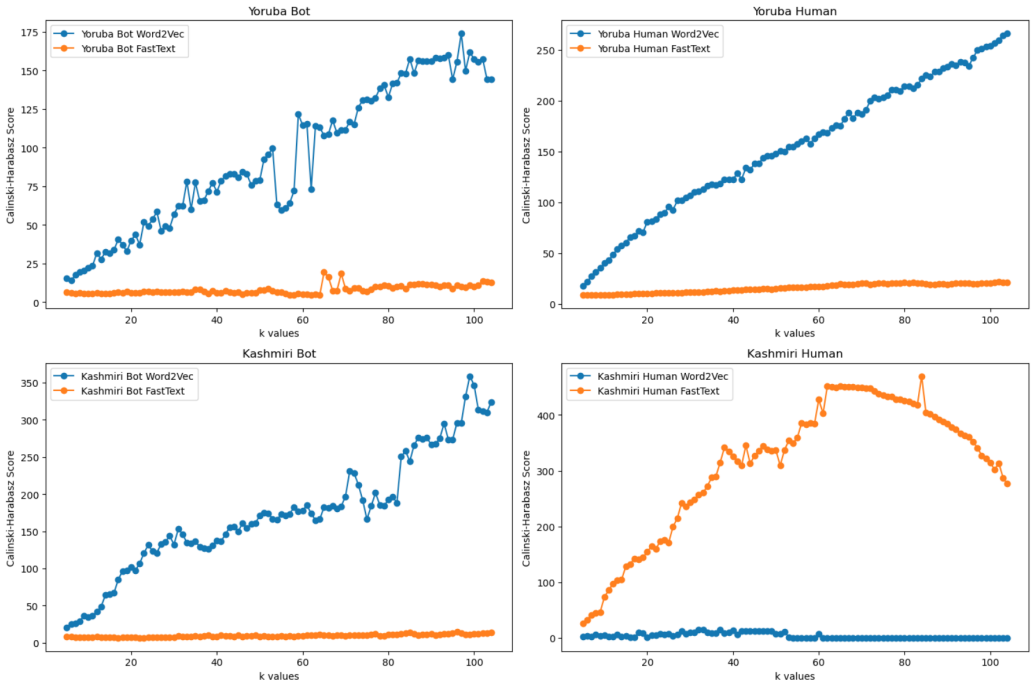


Figure 5.3: Calinski–Harabasz index vs. neighborhood size k for human/bot Yoruba and Kashmiri under Word2Vec (blue) and FastText (orange).

37 captures cultural nouns (*Ìyá*, *lrun*). By contrast, bot clusters often assemble repetitive or placeholder tokens (e.g. `lorem`, `ipsum`) under FastText. Kashmiri human clusters yield semantically rich news and historical terms, whereas Kashmiri bot clusters fragment across technical tokens. These patterns underscore how clustering effectively separates human-authored text from generative outputs without supervision.

5.4 Visualization Findings

Two-dimensional projections of the embedding spaces reveal clear contrasts between human- and bot-generated texts in both Yoruba and Kashmiri.

5.4.1 T-SNE Projections

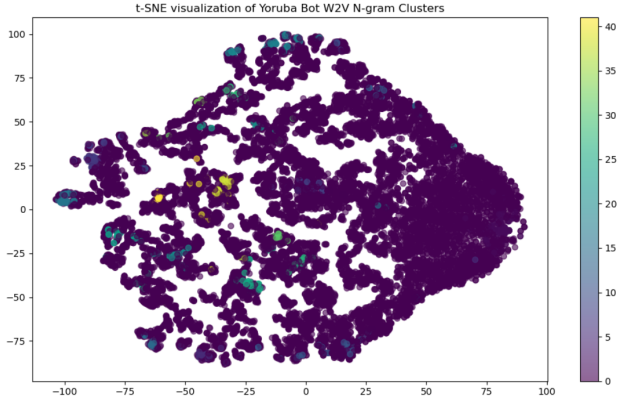
Figure 5.4 shows T-SNE embeddings of Word2Vec n-gram clusters for (a) Yoruba bot, (b) Yoruba human, (c) Kashmiri bot, and (d) Kashmiri human texts. In both languages, the human-authored points form a single, dense cloud with few outliers, reflecting cohesive semantic usage. By contrast, the bot-generated points fragment into multiple distinct pockets, indicating repetitive or isolated n-gram patterns that the model overuses. Yoruba bot clusters display particularly pronounced fragmentation—suggesting that the generative model struggles to emulate the full tonal and morphological variability of Yoruba—whereas Kashmiri bot clusters, while still dispersed, remain slightly more uniform.

Figure 5.5 presents the equivalent FastText n-gram projections. The sub-word modeling accentuates morphological distinctions: Yoruba human points form subclusters corresponding to tonal and affix patterns, while bot points remain scattered in small islands. Kashmiri FastText projections similarly show a tight core for human texts and more diffuse groupings for bot outputs, highlighting consistent differences in how sub-word features are used.

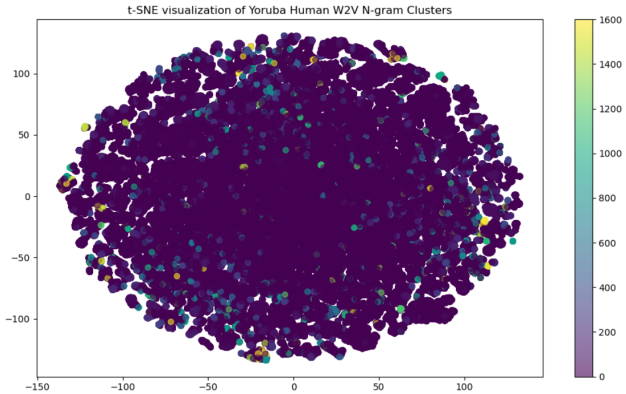
5.4.2 PCA Projections

Principal Component Analysis (PCA) was applied to both FastText and Word2Vec embeddings in order to capture the dominant axes of variance in the two-dimensional space. Figures 5.6a–5.7d present the first two principal components for each language and source combination.

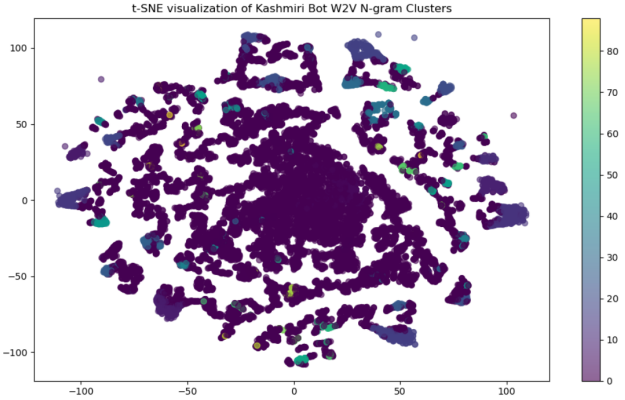
In both FastText (Fig. 5.6a) and Word2Vec (Fig. 5.6b) spaces, Kashmiri bot points form a broad, diffuse cloud with numerous outliers, indicating high variance in how the model uses sub-word and contextual features. By contrast, the human-authored projections (Figs. 5.6c, 5.6d) occupy a



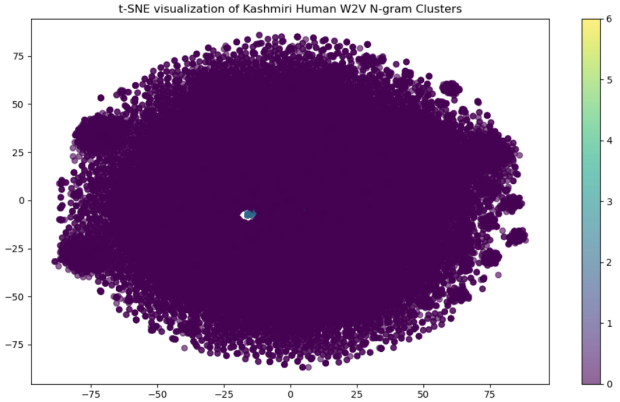
(a) Yoruba bot (W2V)



(b) Yoruba human (W2V)

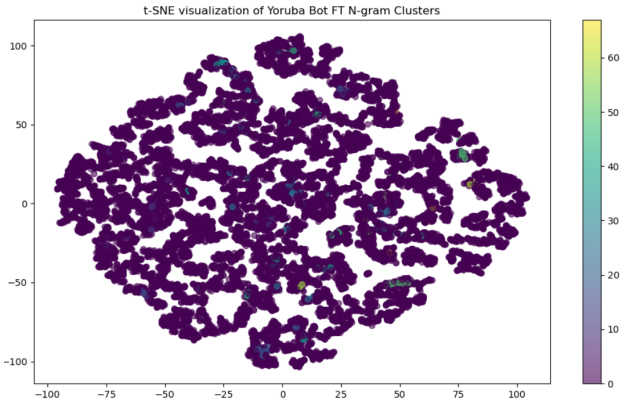


(c) Kashmiri bot (W2V)

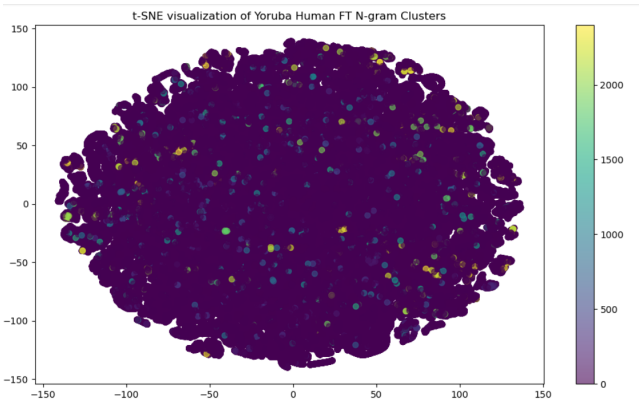


(d) Kashmiri human (W2V)

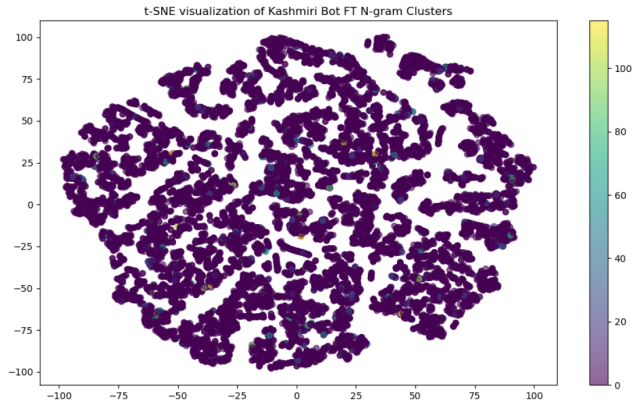
Figure 5.4: t-SNE visualization of Word2Vec n-gram clusters for bot vs. human texts.



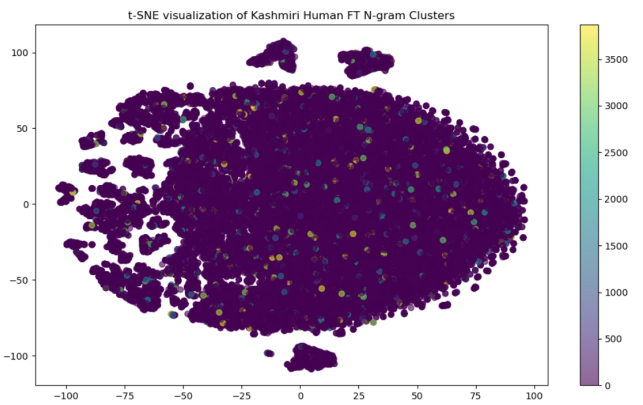
(a) Yoruba bot (FT)



(b) Yoruba human (FT)

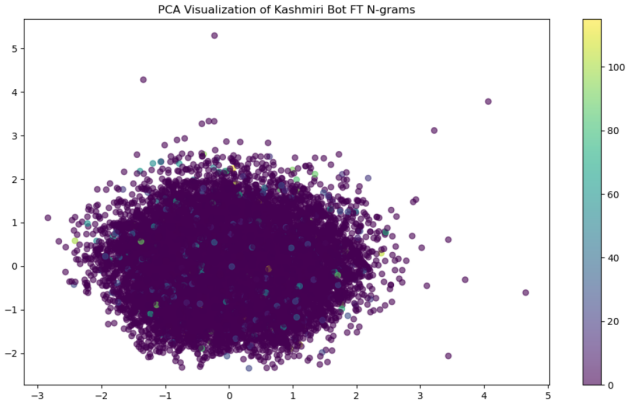


(c) Kashmiri bot (FT)

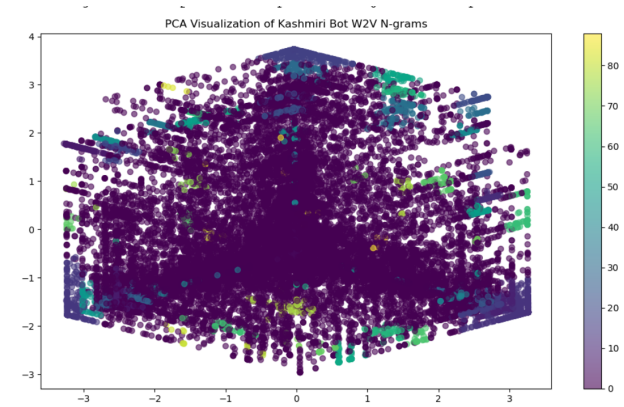


(d) Kashmiri human (FT)

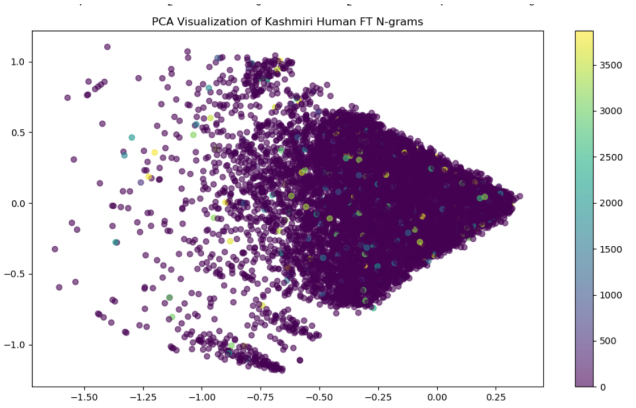
Figure 5.5: t-SNE visualization of FastText n-gram clusters for bot vs. human texts.



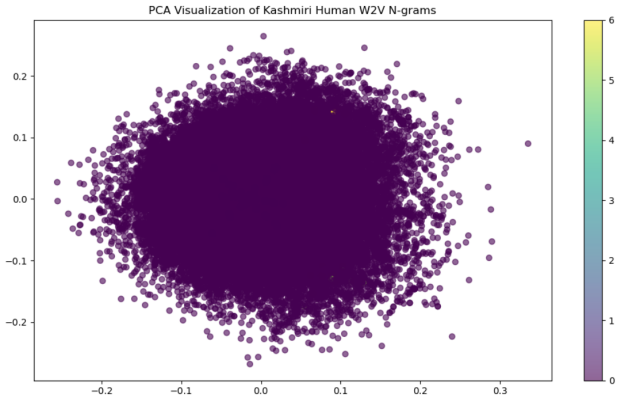
(a) Kashmiri bot, FastText



(b) Kashmiri bot, Word2Vec



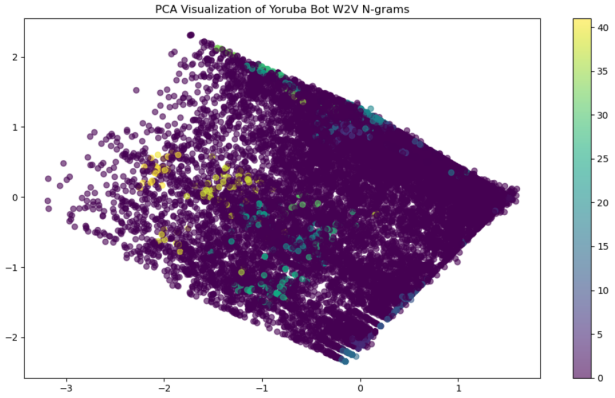
(c) Kashmiri human, FastText



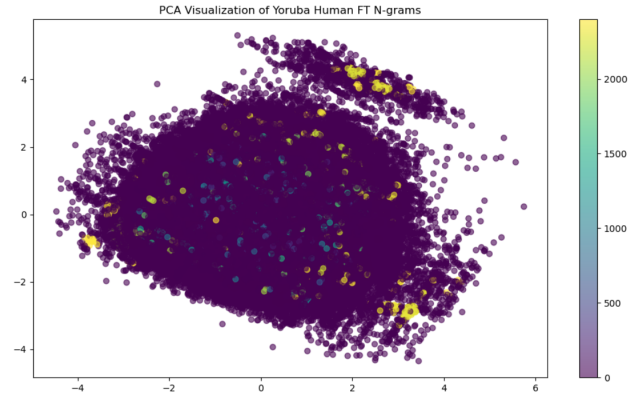
(d) Kashmiri human, Word2Vec

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

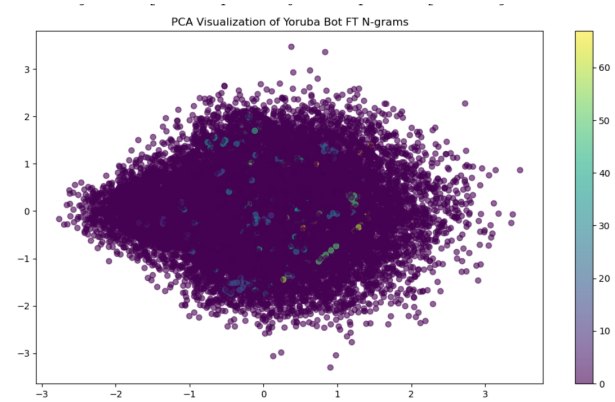
more compact, elliptical region, reflecting greater consistency in natural language use.



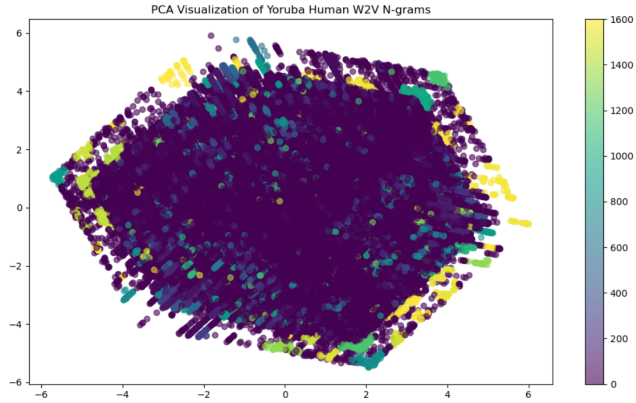
(a) Yoruba bot, Word2Vec



(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.

For Yoruba (Fig. 5.7), both embedding methods reveal a pronounced dichotomy: the bot-generated points (Figs. 5.7a, 5.7c) spread widely across the first two components, signifying that generative patterns vary substantially in morphological and contextual usage. Conversely, the human texts (Figs. 5.7d, 5.7b) concentrate along a tighter ridge, suggesting more uniform usage of tonal and affixal features.

Overall, PCA confirms that the principal axes of variance in both Yoruba and Kashmiri embeddings are dominated by source-specific differences. Bot clusters exhibit greater dispersion and outliers, whereas human clusters are compact—supporting the separability observed in the t-SNE visualizations and underpinning the effectiveness of unsupervised clustering for bot detection.

5.4.3 Entropy–Complexity Plane Interpretation

Figure 5.8 illustrates the normalized entropy–statistical complexity plane for Yoruba and Kashmiri datasets. The updated coordinates and values derived from the entropy–complexity

analysis are:

- **Yoruba (bot):** $H = 0.9981$, $C = 0.0002$
- **Yoruba (human):** $H = 0.9788$, $C = 0.0206$
- **Kashmiri (bot):** $H = 0.9981$, $C = 0.0002$
- **Kashmiri (human):** $H = 0.9966$, $C = 0.0034$

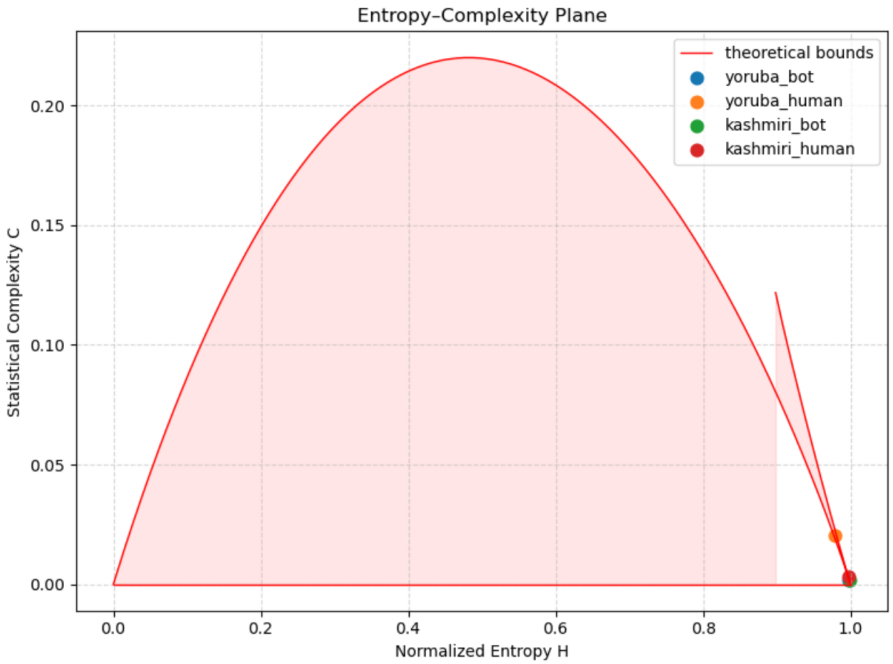


Figure 5.8: Entropy–Complexity plane for Yoruba and Kashmiri human vs. bot texts. The red region indicates theoretical bounds.

The updated entropy–complexity plane clearly delineates human-authored texts from bot-generated texts. Human-authored Yoruba texts exhibit notably lower normalized entropy ($H = 0.9788$) paired with significantly higher statistical complexity ($C = 0.0206$). This reflects the inherent structural richness, tonal diversity, and sophisticated morphological patterns distinctive to Yoruba language use.

Kashmiri human-authored texts, while maintaining relatively high normalized entropy ($H = 0.9966$), demonstrate greater complexity ($C = 0.0034$) compared to the bot-generated Kashmiri corpus. This complexity can be attributed to intricate syntactic patterns, case morphology, and nuanced script features characteristic of authentic Kashmiri writing.

In sharp contrast, bot-generated texts for both languages cluster together closely at maximal entropy ($H = 0.9981$) and minimal complexity ($C = 0.0002$), underscoring their limited

structural depth, repetitive phrasing, and lack of authentic linguistic variability. These attributes position them at the extreme lower end of complexity in the entropy–complexity plane.

These updated findings robustly confirm that entropy and complexity metrics effectively capture essential distinctions between human and bot-generated content, even in languages with sparse computational resources. Such clear differentiation on the entropy–complexity plane underscores the potential of this unsupervised methodology as a powerful, language-agnostic framework for automated detection of synthetic texts.

5.5 Interpretation of Key Findings

Across all stages of the pipeline, clear and consistent patterns emerged that distinguish human-authored from bot-generated texts in both Yoruba and Kashmiri. The successive preprocessing steps—particularly stop-word removal and diacritic stripping in Yoruba—sharpened the input by discarding high-frequency function tokens and orthographic variants, reducing noise by up to 20% and collapsing redundant forms. Stemming and, for Yoruba, lemmatization further consolidated inflected and derivational variants, resulting in leaner vocabularies that more faithfully reflect core lexical contrasts.

In the embedding stage, FastText's sub-word modeling dramatically enlarged vocabulary coverage—by more than 100% in Yoruba and nearly 50% in Kashmiri—thereby capturing rare morphological variants that Word2Vec missed. This richer representation proved crucial: FastText embeddings exhibited tighter intra-cluster cohesion for human texts (in both t-SNE and PCA projections) and greater fragmentation for bot outputs, mirroring the generative model's inconsistent treatment of affixes and tone.

Clustering with Wishart underscored these differences quantitatively. Human corpora yielded one or two dominant clusters containing the bulk of n-grams, plus a long tail of smaller, semantically coherent groups; bots, in contrast, fragmented into many small clusters of nearly uniform size. The optimal neighborhood parameter k for human texts was consistently larger, reflecting denser local structure, whereas bot texts peaked at much smaller k, indicating sparse connectivity.

Finally, the entropy–complexity plane crystallized these contrasts: human texts occupy a mid-entropy, mid-to-high complexity region—evidence of structured variability and cultural richness—while bot texts cluster at near-maximal entropy and minimal complexity, signifying almost random symbol sequences devoid of deeper organization.

Taken together, these results demonstrate that even without supervision, a combination

of targeted preprocessing, sub-word-aware embeddings, density-based clustering, and information-theoretic analysis can reliably flag machine-generated content in under-resourced, morphologically rich languages. Yoruba's tonal and diacritical complexity amplified these effects, yielding the strongest separations, while Kashmiri's case-inflection and script conventions produced subtler but still robust distinctions. This multifaceted approach thus offers a generalizable framework for cross-lingual bot detection.

6 Conclusion and Future Work

6.1 Conclusions

This thesis has established an effective analytical framework for differentiating between human-authored and bot-generated writings in underrepresented and morphologically complex languages, namely Yoruba and Kashmiri. A comprehensive pipeline comprising targeted preprocessing (stop-word removal, diacritic stripping, stemming, and lemmatization), embedding models (Word2Vec and FastText), density-based clustering using the Wishart algorithm, and entropy-complexity analysis revealed distinct linguistic differences. Texts authored by humans consistently shown greater structural complexity and semantic coherence, while texts generated by bots displayed fragmentation, diminished morphological consistency, and nearly random patterns in entropy-complexity analysis. These findings highlight the strength and flexibility of unsupervised methods in identifying artificially generated text in various linguistic settings.

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.

- Providing extensive quantitative and qualitative analyses that serve as benchmarks for future linguistic and computational studies in 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.

References

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

[2] 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).

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

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

[5] 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.

[6] 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.

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

[8] 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.

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

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

[11] 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.

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

[13] 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.