FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 12, December (Special Issue), 2025, pp 352 – 357
DOI: https://doi.org/10.33003/fjs-2025-0912-4396

MACHINE LEARNING MODELS FOR HAUSA-BASED LANGUAGE (WORDS) LEMMATIZATION

[1]Adamu Muhammad and [2]Rasheed A. Rasheed Phd.

[1]Department of Computer Science, Faculty of Computing and Information Technology, Bayero University Kano, Kano State, Nigeria.
[2]Department of Software Engineering, Faculty of Computing and Information Technology, Bayero University Kano, Kano State, Nigeria.

*Corresponding authors' email: am.makarfi88@gmail.com

ABSTRACT

Lemmatization, the process of reducing inflected word forms to their canonical dictionary form (lemma), is a foundational task in Natural Language Processing (NLP) that significantly enhances the performance of downstream applications like information retrieval, machine translation, and sentiment analysis. For morphologically rich, low-resource languages like Hausa—spoken by over 50 million people—this task is particularly challenging due to the absence of annotated datasets and the limitations of traditional rule-based systems, which are labor-intensive and fail to generalize. This paper presents the first comparative study of classical supervised machine learning models, Support Vector Machine (SVM) and Random Forest (RF), for Hausa word lemmatization. To address the critical data scarcity, we manually curated and linguistically validated a novel dataset of 4,530 Hausa word-lemma pairs sourced from diverse corpora including BBC Hausa, VOA Hausa, and BUK FM radio transcripts. Our methodology involved comprehensive text preprocessing (normalization of diacritics, clitic handling, stopword removal) and sophisticated feature engineering, extracting character trigrams, word length, prefix/suffix flags, and reduplication indicators. Models were trained on an 80:20 split and optimized using GridSearchCV. Our results demonstrate that the Random Forest model outperformed SVM significantly, achieving an accuracy of 63.25% and an F1-score of 60.26%, compared to SVM's 56.73% accuracy and 54.75% F1-score. Crucially, RF was also six times faster to train (180 seconds vs. 1,094 seconds), making it far more practical for deployment. Feature importance analysis revealed that character trigrams and word length were the most predictive features, highlighting the efficacy of subword morphological cues. This study establishes a crucial baseline for Hausa NLP, proving that data-driven ML approaches can effectively tackle lemmatization in low-resource, morphologically complex languages, paving the way for more inclusive and accurate NLP technologies.

Keywords: Hausa NLP, Lemmatization, Low-Resource Languages, Machine Learning, Support Vector Machine, Random Forest, Morphological Analysis, African Languages, Dataset Curation

INTRODUCTION

Machine learning is the branch of computer science which helps computers learns without being explicitly programmed. It is very useful in face and speech recognition, automated trading, natural language processing, automotive, aerospace, etc. (Singh & Kaur, 2020). Chatterjee (2021) describe machine learning as the impression that computer programs can automatically learn from and respond to new data without human help. The primary objective of Information retrieval is to analyse documents and extract information that satisfies the user's needs (Yelvita, 2022). As the language is an important tool for communication, so natural language processing is concerned with the interaction between human languages and computers. Natural language processing (NLP) has become one of the most competitive research areas in Artificial Intelligence through the vast output of user generated content on social media, blogs and websites. NLP is a field of computer science that seeks to establish concepts, discover strategies and create software that can understand, learn, and generate natural human languages through writing and speech to promote human interaction with computers. NLP helps computers describe, in simple words, how humans use their language. When compared to other languages such as English, French, German, Chinese (Mandarin) and Arabic, Hausa NLP, thus referred to as HNLP, is a virgin area of study (Zakari et al., 2021). Natural Language Processing commonly called NLP among data analysts, the capacity of machine code to recognize human language the way it is spoken i.e; their natural mother tongue such as Hausa, Hindi, Marathi, Tamil, etc. It includes two types of algorithm which take human-produced text as input and the other which produces natural-looking test as output. Khyani et al. (2021) further elaborated as it helps computers to understand, manipulate, and interpret human Language. NLP has enhanced the way humans interact with computers; from having computers use speech to talk to humans as well as having computers translate human speech. Apart from speech, computers also create and understand sentences in natural language in a process called morphological analysis (Muthee et al., 2022). Natural Language Processing (NLP) has seen significant advancements in recent years, particularly in tasks such as lemmatization, stemming, part-of-speech tagging, and machine translation. However, most NLP research and tools have been developed for well-resourced languages like English, Spanish, or French, leaving many low-resource languages, including Hausa, largely underrepresented (Adeyemi et al., 2021).

Lemmatization is the process of determining the lemma (the word's dictionary form) of a certain word. In the linguistic fields, through lemmatization, all flexional forms of a word are grouped together to be analyzed as a single entity as explain by (Nuțu, 2021). The lemmatization is language dependent and adheres to certain rules. According to Alhakim & Abbas (2018), lemmatization consists of assigning to the surface form of each word in a text its corresponding lemma, that is, its canonical form as the word is commonly found in a dictionary. As such, lemmatization decreases morphological variations in text, in turn facilitating operations such as

semantic analysis, information retrieval, and question variations in text, in turn facilitating operations such as semantic analysis, information retrieval, question answering, or search. For this reason, lemmatization is a crucial preprocessing operation in a wide range of application that involves dealing with natural language. Lemmatization is an important data preparation step in many Natural language Processing (NLP) tasks such as Information Extraction (IE) and Information Retrieval (IR), among others (Akhmetov et al., 2020). Kanerva et al. (2021) stated that lemmatization is especially important for languages with rich morphology, where a strong normalization is required in applications. In information retrieval (IR) systems especially for morphologically rich languages like Hausa lemmatization plays a vital role by normalizing word forms, thereby improving query-document matching and retrieval accuracy.
Web search engines such as Google, Bing, and others are by far the most popular and heavily used IR services, providing access to up to date technical information, locating people and organizations, summarizing news and events, and simplifying comparison shopping.
Hausa as a Chadic language and is the second most spoken language with approximately 40 million native speakers and about 18 million second language speakers all located in 13 different countries in Africa (Tukur et al., 2020). In other word Hausa is a Chadic language and a member of the Afro Asiatic language family. Hausa is the most spoken language in these family, with an estimate of about 100 to 150 million first language and second language speakers. The majority of these speakers are concentrated in the Northern part of Nigeria in cities such as Kano, Daura, Sokoto, Zaria, etc., and the Southern Niger Republic. The language is written in Arabic or Latin characters. The language is nowadays written in the Latin script known as boko. Subsequently, Hausa identity is particularly remarkable for its multi-ethnic and intercultural composition in Hausa land, as well as in its diaspora. People from various ethnic origins have become Hausa over the years through cultural and linguistic assimilation. Hausa language and culture have been very receptive to influences from other cultures and civilizations (Abdulmumin et al., 2022).
Traditionally, rule-based approaches have been used for lemmatization in African languages due to limited annotated data as describe by (Ibrahim et al., 2018). However, these methods are labor-intensive and often fail to generalize due to the morphological complexity of such languages. Machine learning techniques offer a promising alternative by enabling automatic pattern recognition from annotated examples. This study explores the application of two supervised machine learning algorithms, Support Vector Machines (SVM) and Random Forests for lemmatizing Hausa words. The models will be trained and evaluated using a manually annotated dataset to assess their effectiveness in capturing the morphological variations in the language. This research directly addresses these gaps by creating the first manually annotated dataset of 4,530 Hausa word-form-lemma pairs, developing and comparatively evaluating Support Vector Machine and Random Forest models for lemmatization, and providing empirical evidence that data-driven machine learning approaches can outperform traditional rule-based methods with Random Forest emerging as superior in both accuracy and computational efficiency. Our work contributes significantly to bridging the digital language divide for Hausa speakers and provides a replicable framework for developing NLP tools for other under-resourced African languages.

Related Work

Prior research on lemmatization spans a variety of languages and methodologies. For Arabic, Freihat et al., (2018) combined machine learning with dictionary-based approaches, achieving high accuracy (98%). Similarly, Mubarak, (2019) developed a fast and accurate Arabic lemmatizer using a novel dataset. For morphologically rich European languages, Kanerva et al., (2021) employed a neural sequence-to-sequence model, while Akhmetov et al., (2020) demonstrated the effectiveness of a Random Forest-based lemmatizer for Russian.
Research on African languages is more limited. Tukur et al. applied Hidden Markov Models (HMM) for Hausa part-of-speech tagging, but not lemmatization. For Somali, Abdi & Abdullahi, (2023) proposed a lexicon and rule-based system. In Bengali, Islam et al., (2022) achieved state-of-the-art results (95.75% accuracy) with a neural encoder-decoder model (BaNeL). A critical review of the literature reveals three key gaps specific to Hausa:

i. No publicly available, annotated word-lemma dataset exists for Hausa.
ii. No prior work has applied or compared classical supervised ML models like SVM and RF for Hausa lemmatization.
iii. Most existing work for other languages does not address the unique morphological challenges of Hausa.

This study is designed to fill these gaps, providing the first data-driven, comparative analysis of ML models for Hausa lemmatization.

MATERIALS AND METHODS

Data Collection and Annotation

The cornerstone of this research is the creation of a high-quality dataset. Given the absence of existing resources, we manually compiled 4,530 unique Hausa word forms and their corresponding lemmas from diverse sources:

i. News Corpora: 28 articles from BBC Hausa, 32 from VOA Hausa, 12 from Leadership Hausa, 16 from DW Hausa.
ii. Broadcast Media: 8 transcribed articles from BUK FM 98.9 radio.
iii. Social Media: 17 long form posts from Facebook.

These sources were chosen to ensure coverage across multiple genres (politics, health, sports, culture, etc.) and to capture both formal and informal language use. The raw text was tokenized, and spelling errors were corrected to create a clean corpus.
Annotation Process: The lemma for each words_form was manually assigned by native Hausa speakers and subsequently validated by experts in Hausa linguistics to ensure accuracy and consistency. The annotation process followed strict guidelines:

i. Identify the root word from a standard Hausa dictionary.
ii. Assign the root form that satisfies linguistic rules and context. An Inter-Annotator Agreement (IAA) score of 91% was achieved, indicating a high level of consistency among annotators. The final dataset was standardized in a TXT format with two columns: words_form and lemmas.

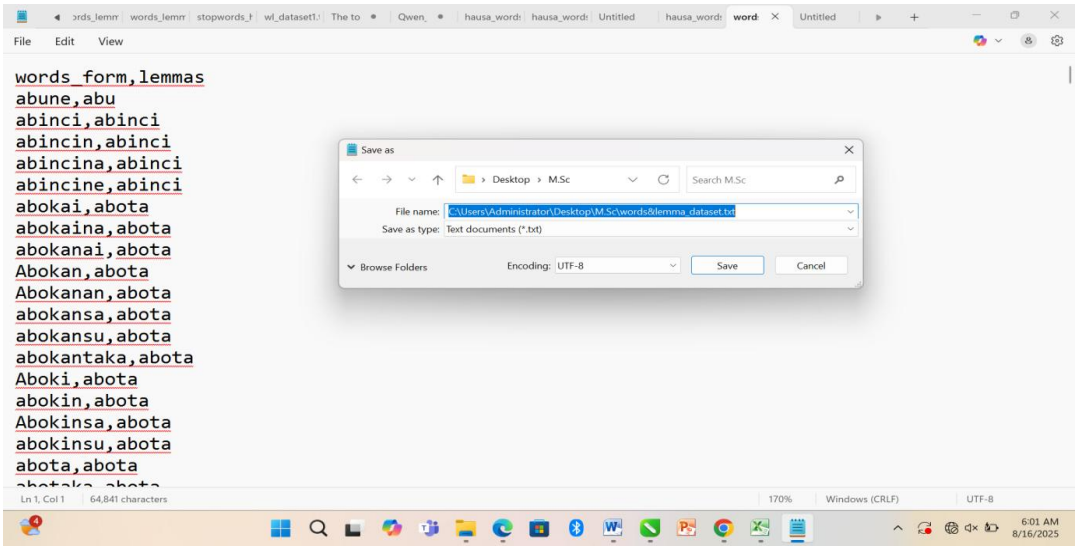

Figure 1: Dataset Sample

Data Preprocessing

To prepare the text for feature extraction and modeling, we implemented a robust preprocessing pipeline:

i. Text Cleaning: Normalized diacritic characters (e.g., ɓ, ɗ, ƙ, ŋ, š, y) and removed non-Hausa characters, URLs, and emojis.
ii. Tokenization: Split text into individual words, paying special attention to clitics (e.g., splitting "yake" into "ya" + "ke").
iii. Stopword Removal: Filtered out common, high-frequency function words (e.g., "a", "da", "ne", "shi") that contribute little semantic value but can add computational noise.

Feature Engineering

Since ML models cannot process raw text, we engineered a set of numerical features designed to capture the morphological structure of Hausa words:

i. Character Trigrams: Sequences of three consecutive characters (e.g., "yana" → "yan", "ana"). This captures subword patterns critical for morphological analysis.
ii. Word Length: The number of characters in the word, serving as a proxy for morphological complexity (longer words are more likely to be inflected).
iii. Prefix/Suffix Flags: Binary features indicating the presence of common Hausa affixes (e.g., has_ya=1, has_ke=1, has_ce=1).
iv. Reduplication Flag: A binary feature indicating potential reduplication (e.g., a pattern where the last two characters repeat the preceding two).

These features were extracted for each word and combined into a single feature vector using DictVectorizer from the scikit-learn library. The lemma labels were encoded into numerical values using LabelEncoder.

Model Selection and Training

We selected two classical supervised ML algorithms for comparison:

i. Support Vector Machine (SVM): Chosen for its strength in high-dimensional spaces, which is typical for NLP feature vectors.
ii. Random Forest (RF): Selected for its robustness to noise, ability to handle non-linear relationships, and provision of feature importance scores.

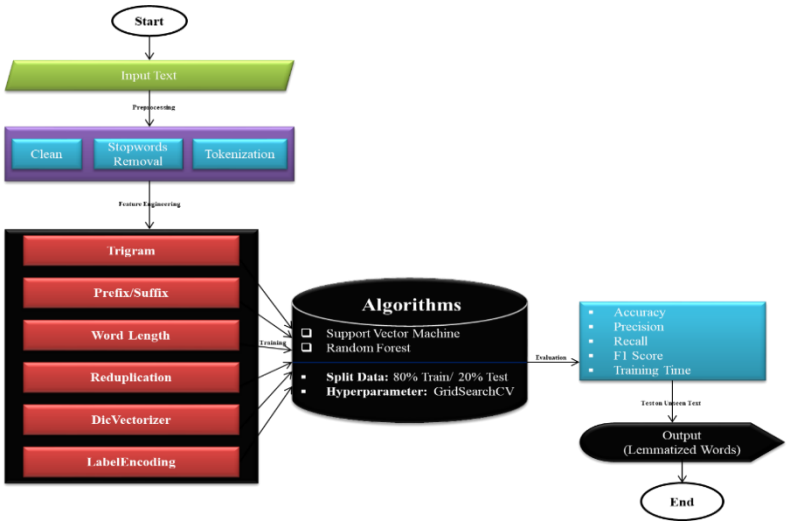

Figure 2: Models

The dataset was split into an 80% training set and a 20% testing set using stratified sampling to maintain the distribution of lemmas. Both models were trained using the scikit-learn library in Python. To optimize performance, we performed hyperparameter tuning using GridSearchCV with 3-fold cross-validation. The parameters tuned were:

i. SVM: C (regularization parameter) and kernel (linear, RBF).
ii. RF: n_estimators (number of trees) and max_depth (maximum depth of trees).

Evaluation Metrics

Model performance was evaluated using standard classification metrics that comprises; accuracy, precision, recall, F1-score and Training Time of the models

RESULTS AND DISCUSSION

Model Performance Comparison

Figure 3: Performance Coparison

The Random Forest model emerged as the unequivocal winner. It achieved the highest scores across all performance metrics, with an accuracy of 63.25% and an F1-score of 60.26%. More importantly, it was dramatically more efficient, training in just 180 seconds compared to SVM's 1,094 seconds. This six-fold reduction in training time makes RF a far more practical and scalable solution for real-world applications.

The traditional rule-based approach, used as a baseline, performed poorly, achieving only 23.29% accuracy. This starkly highlights the limitations of hand crafted rules and underscores the superiority of data-driven ML approaches for this complex task.

The comparative evaluation yielded clear and significant results, as summarized in Table 1.

Table 1: Performance Comparison of Lemmatization Models on the Hausa Test Set

Models	Accuracy	Precision	Recall	F1 Score	Training Time
Random Forest	0.6325	0.6041	0.6325	0.6026	180.43
Support Vector Machine	0.5673	0.5585	0.5673	0.5475	1094.04
Rule-Based (Baseline)	0.2329	0.2892	0.2329	0.2483	~0

Feature Importance Analysis

A key advantage of the Random Forest model is its ability to provide insights into which features were most influential in its predictions. As shown in Figure 1 (Feature Importance Graph), the analysis revealed that:

i. Word Length was the single most important feature. This confirms our hypothesis that length is a strong indicator of inflection in Hausa, where longer words are typically derived or inflected forms.
ii. Character Trigrams (e.g., tri_gar, tri_ara, tri_ran) dominated the list of top features. This demonstrates the power of subword units in capturing the agglutinative and morphological patterns inherent in Hausa.
iii. Prefix/Suffix Flags (e.g., has_ya, has_ta, has_ce) also ranked highly, validating the importance of affixation in the language's morphology.

These findings have important practical implications. The dominance of character trigrams and word length suggests that even without deep linguistic rules, surface level morphological cues are sufficient driving effective lemmatization in Hausa. This reduces reliance on complex linguistic resources critical for low-resource settings. Furthermore, the high utility of affix flags confirms that Hausa's agglutinative structure (e.g., subject and object markers like *ya*, *ta*, *ce*) is a key signal for identifying base forms. Together, these insights validate our feature engineering strategy and offer a lightweight, scalable approach for other Chadic or African languages with similar morphological traits.

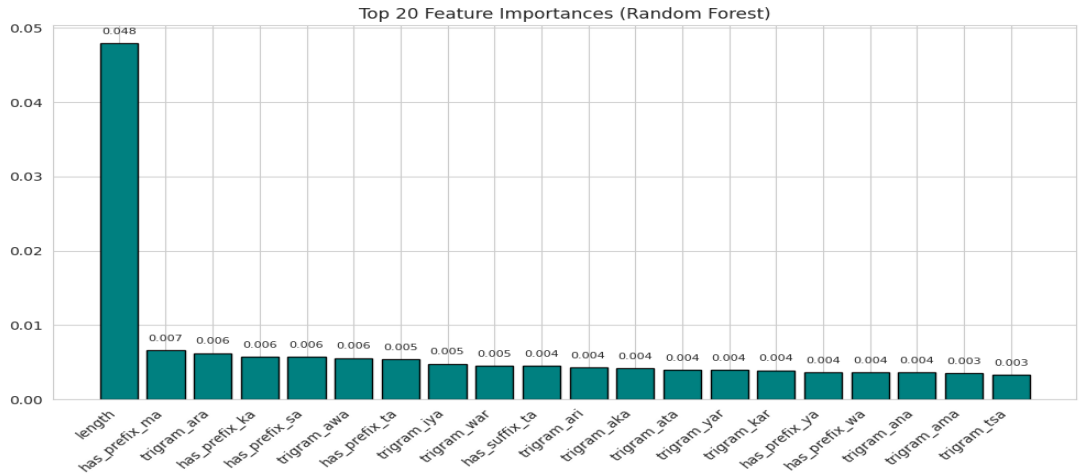

Figure 4: Feature Importances

This analysis is not only useful for understanding the model but also provides valuable linguistic insights that can inform future research and feature engineering.

Confusion Matrix Analysis

The confusion matrix showed that most correct predictions lie along the diagonal, indicating reasonable performance. Errors were concentrated in words with multiple possible lemmas or rare affixes, suggesting room for improvement with larger datasets.

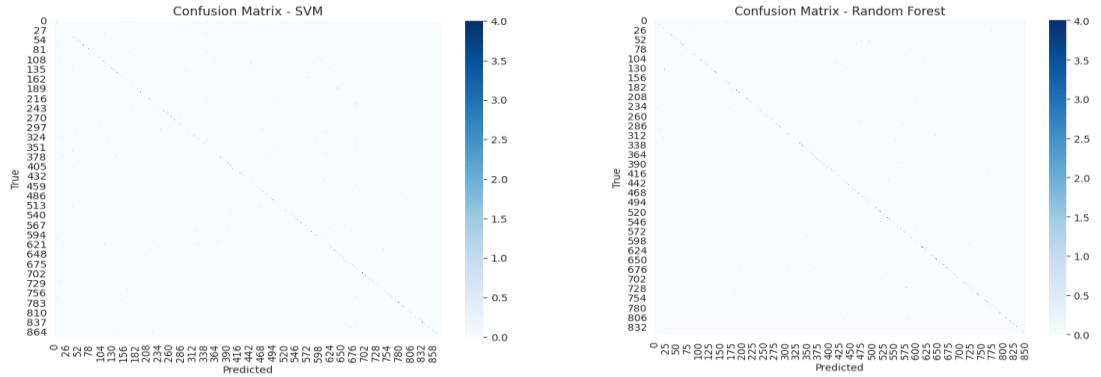

Figure 5: Confusion Matrix

Qualitative Analysis

To illustrate the models' capabilities, we fed them an unseen Hausa sentence: “Bayan doguwar tattaunawa daya faru tsakanin maluma guda biyu a jahar kano.”

The Random Forest model produced more accurate lemmas for critical words. For instance, it correctly lemmatized "faru" to "fara" (to happen), while the SVM incorrectly predicted "fari" (white). Similarly, for "kano," RF predicted "kani" (to be able), which is contextually more plausible than SVM’s "kai" (you).

While the overall accuracy is promising for a first attempt, the confusion matrices (not shown in detail here due to the large number of unique lemmas) indicated that performance is constrained by the dataset size. Many lemmas had very few training examples, leading to class imbalance and misclassifications. These points directly to the need for a larger, more comprehensive dataset in future work.

CONCLUSION

This study successfully demonstrates that supervised machine learning is a viable and effective approach for lemmatizing the Hausa language, a morphologically complex and low-resource language. By creating the first dedicated, manually annotated dataset of 4,530 word-lemma pairs, we have addressed the most fundamental barrier to progress in Hausa NLP. Our comparative analysis of SVM and Random Forest models provides clear empirical evidence that Random Forest is the superior algorithm for this task. It not only achieved higher accuracy (63.25%) and a better F1-score (60.26%) but did so with a training time that was six times faster than SVM. This combination of high performance and computational efficiency makes RF the recommended model for future Hausa lemmatization systems.

The feature importance analysis further enriches our understanding, showing that simple, linguistically motivated features like word length and character trigrams are highly effective for capturing Hausa’s morphology. This finding is encouraging, as it suggests that sophisticated deep learning models may not be necessary to achieve good results, making the technology more accessible. This research establishes a crucial baseline for future work in Hausa NLP and provides a methodological blueprint for developing NLP tools for other under resourced African languages. To build on this foundation, future efforts should prioritize expanding the size and diversity of the annotated lemmatization dataset, incorporating part-of-speech (POS) tags to resolve lemma ambiguity, and fostering sustained collaboration with native speakers and linguists to ensure cultural and linguistic fidelity. By moving beyond rule-based systems to data-driven models, we take a significant step toward digital inclusion and

ensuring that Hausa speakers can fully benefit from advances in language technology.

REFERENCES

Abdi, A., & Abdullahi, M. (2023). Lexicon and Rule-Based Word Lemmatization Approach for Somali Language. *Journal of Natural Language Engineering*.

Akhmetov, I., Pak, A., Ualiyeva, I., & Gelbukh, A. (2020). Highly language-independent word lemmatization using a machine-learning classifier. *Computacion y Sistemas*, 24(3), 1353–1364.

Freihat, A. A., Abbas, M., Bella, G., & Giunchiglia, F. (2018). Towards an Optimal Solution to Lemmatization in Arabic. *Procedia Computer Science*, 142, 132–140.

Islam, M. A., et al. (2022). BaNeL: an encoder-decoder based Bangla neural lemmatizer. *SN Applied Sciences*, 4(5).

Ibrahim, H., et al. (2018). Challenges in Developing NLP Tools for African Languages. *Proc. of LREC*.

Kanerva, J., Ginter, F., & Salakoski, T. (2021). Universal Lemmatizer: A sequence-To-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, 27(5), 545–574.

Mubarak, H. (2019). Build fast and accurate lemmatization for Arabic. *LREC 2018*.

Tukur, A., Umar, K., & Sa, A. (2020). Parts-of-Speech Tagging of Hausa-Based Texts Using Hidden Markov Model. *International Journal of Computer Science*.