Vol. 07, No. SI♣, pp. 37 –45 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Part-of-Speech Tagging-Based Document Clustering for Kurdish Corpora

Toreen Dilshad Masoud[1,*], Ismael Ali Ali[2]

[1]*Department of Information Technology, Technical College of Informatics, Akre University for Applied Sciences, Akre, Kurdistan Region, Iraq, toreen.dilshad@auas.edu.ac*
[2]*Jazari Research Center, Research Center, University of Zakho, Zakho, Kurdistan Region, Iraq, ismael.ali@uoz.edy.krd*

**Correspondence: toreen.dilshad@auas.edu.ac*

Abstract

This study investigates the Natural Language Processing (NLP) challenges of unsupervised clustering for Kurdish documents, focusing on the role of Part-of-Speech (POS) tagging in improving clustering performance. Due to the linguistic complexity of Kurdish and the scarcity of annotated corpora, the proposed method used a TF-IDF (Term Frequency–Inverse Document Frequency) matrix with K-Means clustering, applied POS tagging to the Badini Kurdish corpus. POS tagging is important for capturing the syntactic and grammatical structures of text. The experiments were conducted using the UOZBDN corpus, which includes 231 documents distributed across five categories. To evaluate the impact of POS tag. Additionally, compared the clustering performance using POS tagging and No POS tags of the same corpus. One of the main challenges in this research is the absence of prior studies that apply document clustering techniques to Kurdish corpora. Therefore, there is limited prior work available for direct comparison with the results obtained in this study. The results showed that incorporating POS tags, particularly a carefully selected subset of 22 key of POS categories, significantly improved clustering performance. The proposed approach achieved a Purity score of 0.9714, an NMI score of 0.9477, and a Silhouette score of 0.2403. demonstrated that POS tagging significantly enhanced clustering quality and highlighted the importance of POS represented in Badini Kurdish corpus.

Keywords: *Natural Language Processing, Document Clustering, Kurdish Language Processing, TF-IDF, Part-of-Speech Tagging*
Received: October 04th, 2025 /Revised: March 25th, 2026 / Accepted: April 15th, 2026 / Online: April 19th, 2026

I. INTRODUCTION

Document clustering is a crucial task in Natural Language Processing (NLP), enabling the automatic organization of large text corpora into meaningful groups for faster information retrieval processes by clustering the corpus documents into topic-based clusters [1]. while keeping the different clusters apart [2].Such data points within each of these clusters are similar to one another but distinct from the rest of the information [3]. K-Means is a well-known unsupervised clustering technique that finds a structure in the provided datasets. The clustering method is widely used for dividing data into groups and is widely used due to its simplicity. Although its applications have been observed in various critical real-world problems [4]. Additionally, preferred simplicity and fast execution, particularly with respect to the number of iterations, in several applications [5].

Document clustering begins with text representation, a process that converts raw text into a numerical format using simple rule-based or statistical approaches [6]. Then, Term Frequency-Inverse Document Frequency (TF-IDF) is used to represent the frequency at which words appear in documents within the corpus. TF-IDF creates term weights based on the frequency of each word in a document. TF scores each feature based on its frequency in each document, and IDF computes the number of unique features that can be found in the dataset [7]. TF-IDF is used in other NLP applications, such as sentiment analysis of text data, which uses the TF-IDF technique for its efficiency and quick computation time [8]. The Part of Speech (POS) tagging, or grammatical tagging, consists of assigning POS tags to each word/token in a given corpus of text. The traditional definitions consider POS tagging a building block for other NLP applications, including named entity recognition, information extraction, spelling correction, text classification, natural language generation, and machine translation. Similar to POS tagging, which can be regarded as an early-stage step for syntactic parsing tasks [9].

doi: 10.38094/jastt7031086

This study quantifies the impact of POS filtering on Kurdish document clustering. this gap was closed by presenting document clustering using selective POS tags to restrict the development of robust tools for the organization and retrieval of Kurdish text. The scarcity of linguistic resources in Kurdish, especially for the Badini dialect. In this work Using a TF-IDF representation and K-Means clustering, investigates various settings of POS tags in the corpus and their impact on the quality of clustering experimental results on a POS-tagged Kurdish corpus of the UOZBDN corpus, to show significant improvements in Purity, Normalized Mutual Information (NMI), and Silhouette Score on a POS-tagged dataset for the Badini dialect of the Kurdish language. The proposed method incorporates POS information into the clustering process, contributing to more accurate evaluation performance in improving document clustering for low-resource dialect.

II. Related Work

Several researchers have explored text clustering using traditional representations and algorithms, as indicated in Table I, and clustering using POS tags, shown in Table II. Key studies include the following:

Mustafa and Jacksi [10] tackled document grouping from a meaning-oriented angle rather than relying on surface-level term counts. Working with a small collection of one hundred movie synopses (drawn from IMDB and Wikipedia) plus the txt_Sentoken and NLTK_Brown corpora, they paired Affinity Propagation for the tiny sets with K-Means for the larger ones, hoping to balance cluster quality against runtime. Texts were encoded with TF–IDF, and while Affinity Propagation reached 36.7 % and 52.5 % purity on NLTK_Brown and txt_Sentoken, respectively (K-Means lagged in these tiny splits), its Silhouette never climbed past 0.03, and it collapsed on bigger inputs. The study is interesting for highlighting a hybrid strategy; however, its semantic signal remains shallow (plain TF–IDF), and Affinity Propagation does not scale, resulting in modest overall accuracy.

Salih and Jacksi [11] revisited the same general question on three large test beds—IMDB reviews (50,000 documents), Reuters-21578 (approximately 10,000), and 20 Newsgroups (20,000). Using TF–IDF and a simple bag-of-words profile, they compared K-Means with Ward's hierarchical method and judged the outcomes through a broad metric suite (Silhouette, Purity, V-Measure, F1, Accuracy, Homogeneity, Completeness, NMI). With basic pre-processing, K-Means achieved a Silhouette of 0.384 and Purity of 0.62 on the Reuters subset, outperforming Ward's on most counts. Weak spots were quickly exposed: minimal text cleaning and the absence of richer representations meant many clusters still blurred topical boundaries. The authors recommend stronger linguistic pre-processing—stemming, stop-word removal—and moving toward semantic vectors (e.g., Word2Vec or BERT) to remedy the shortfall.

Jacksi and Salih [12] assessed two mainstream algorithms—Hierarchical Agglomerative Clustering (HAC) and K-Means—across four heterogeneous sets (IMDB, Wikipedia, 20 Newsgroups, and txt_Sentoken). All documents were mapped into TF–IDF space, then evaluated with Purity, Silhouette, and the Adjusted Rand Index. HAC edged ahead of K-Means on both 20 Newsgroups (Purity 0.106 vs 0.063) and txt_Sentoken (0.611 vs 0.164), but absolute scores were low, highlighting how both algorithms buckle under high-dimensional text. The authors trace this weakness to scalability and sparsity; they suggest first compressing vectors with PCA or t-SNE, or switching to more scalable schemes such as DBSCAN or mini-batch K-Means when handling larger corpora.

Kumbhar and Mhamane [13] Proposed dimensionality reduction methods followed by clustering to cluster common themes in large-scale unstructured textual data. The dataset used was 20-Newsgroups, and the text representation used was TF-IDF. The study applied different combinations of algorithms, such as K-Means, SVD+K-Means, and NMF+K-Means, to achieve better clustering performance. The performance of clustering was assessed based on the accuracy, clustering purity, and visualizations of the approaches, which were best with the Homogeneity: 0.786, Completeness: 0.818, and ARI: 0.876 at K = 2. But that study did not address scalability and computational efficiency, key issues in working with big data. The non-scalable methods, like SVD, can be replaced with relatively more efficient techniques, like Truncated SVD, to reduce the overall computational cost and make it easier to build clustering applications on large-scale datasets.

Perumal and Mathivanan [14] highlighted automatic document clustering and topic identification in the scope of real-time applications using a unique set of text preprocessing and advanced optimization techniques. In the preprocessing stage, executed tokenization, stop-word elimination, and stemming, and then used TF-IDF, Mutual Information (MI), and TextRank as keyword extractors. Used a feature set of terms, mutual information scores, and extracted keywords that were processed using our main contribution, Type-2 Intuitionistic Fuzzy Clustering and Seagull Optimization Algorithm (Type-2 IFCSOA). A comparison was performed between the performance of this approach and other clustering methods, namely FCM, FCM with Particle Swarm Optimization (FCM-PSO), FCM with Genetic Algorithms (FCM-GA), and K-means. PTC-HTM and NPTC-HTM used Precision, Recall, Accuracy, Sensitivity, Purity, and Entropy as evaluation metrics. Through experiment results, it was shown that the proposed method increased accuracy from 0.75 to 0.80. Nevertheless, the Type-2 IFCSOA approach also struggled with scalability and computational efficiency when tested on large-scale datasets, because of its computationally expensive nature.

Saha [15] explored cluster reviews of Amazon products using different types of text embeddings and algorithms, and faced several limitations. Even with K-Means, single-linkage hierarchical clustering, and density-based methods, DBSCAN using BERT and Word2Vec embeddings, clustering quality was typically not great. Based on external validation metrics, which include ARI along with cluster purity, the clustering performance had a poor agreement with true star ratings, with an ARI close to 0, and a cluster purity between 0.60 and 0.67. Hyperparameter sensitivity meant that results varied greatly by small changes to DBSCAN's epsilon or minimum cluster size, leading to a less-stable result. However, high-dimensional embeddings were not well-distributed, leading to problems with

cluster cohesion and separation, and single-linkage hierarchical clustering resulted in chains of clusters. This was also the case for the number of clusters (three only justified by internal validation vs. 5 rating categories), which resulted to be ambiguity. Also, due to the labeling of a huge number of points as noise, internal metrics like silhouette scores may have overestimated clustering performance. The characteristics of the sample and the size of the dataset were not completely described, which limited generalizability. They did not explore higher-level hierarchical or hybrid clustering techniques. In a nutshell, these limitations highlight difficulties in clustering ultra-dimensional document data and the need for more methodological refinement.

Sampaio and Maxcici [16] presented approach as a new method of unsupervised document clustering that integrates text, layout, and visual modality information into multimodal embeddings. The goal of the researchers was to tackle the problem of clustering not only by document category. Used embeddings for documents using various multimodal models. Used classical clustering algorithms based on centroids (K-Means) and density DBSCAN, to analyze these embeddings by clustering them using their previously created embeddings. Tested clustering by performing experiments on five synthetic datasets (invoice datasets and noisy scanned pages dataset) with mixed aspects. Evaluated the clustering results using ARI, NMI, homogeneity, completeness, and silhouette score. The results confirmed that hybrid multimodal models performed compact and powerful clusters, especially for unsupervised document clustering and intelligent document processing.

Rosell [17] introduced part-of-speech tagging for text clustering in Swedish,relied on two corpora: DN, which consists of about 6,400 news articles from Dagens Nyheter, and Occ, which consists of approximately 42,000 short occasional texts. Using K-Means-clustering. Different text representation methods were tried: word form, lemma, lemma + PoS tags, lemma + PoS + gen. Also experimented with other preprocessing methods, including stoplists and confining features to certain word classes (nouns, proper names, adjectives, verbs, and adverbs) that included PoS tags, but this did not improve clustering quality; however, lemmatization yielded particularly strong improvements in results, especially for the Occ dataset. In certain situations, particularly with brief texts, employing a stoplist yielded better results. More importantly, using only nouns and proper names to find the representation returned similar results to those of using all words, but with far fewer features. In general, the results indicated that, although PoS tagging is used in Swedish text clustering, both lemmatization and feature selection give significant gains.

Ahmed, Baharin, and Nohuddin [18] presented research on clustering-based prioritization on translations of the Holy Quran exegesis into English, with five translators (TR1–TR5) of Surah Al-Baqarah. Due to the lack of ground truth datasets, this research used unsupervised learning methods and not supervised classification. We created an extension of three steps; the first stage comprised text extraction (tokenization, PoS tagging, normalization, stemming and stop-word removal) and feature representation with TF-IDF and VSM; the second stage carried out two clustering methods, K-Means (partitioning based) and Agglomerative Hierarchical clustering; and the third stage carried out clustering validation through Silhouette Coefficient (SC), Davies-Bouldin index (DBI) and execution time (ET) with PCA for visualization. With K-Means three out of the five translators had non-matching ranks across all metrics (only ranks (1) & (3) were consistent in all cases) and with Agglomerative clustering each of the metrics provided a unique rank. The findings of the study suggested that, despite the capability of clustering for revealing structural similarities across translations, an advanced approach like Multi-Criteria Decision-Making MCDM should be employed to get an optimal global ranking. In the future, we intend to apply more clustering techniques, increase the number of translators, add more validation metrics, and report machine-based ranking compared to experts in Islamic studies.

Gupta and Joshi [19] proposed a new technique for automatically labeling large, unlabeled real-time Twitter datasets for sentiment analysis. The method proposed here used POS based, n-grams, Twitter-specific, lexicon-based features and added negation modeling with an improved K-means clustering algorithm for grouping tweets into Positive, Negative, or Neutral classes. Non-English tweets were retrieved on the topics of Demonetization, Lockdown, and 9pm9minutes using the Twitter Search API. Sentiment labels were assigned based on manual inspection of each cluster, with the label being propagated to all tweets assessed to belong to that cluster. By accounting for negation and additional features, the increased quality of the clusters (quantified by silhouette score and inertia) is evident when compared to baseline models that used traditional TF-IDF or random K-means. Tweet counts distribution per cluster; e.g. there were 4976 Positive, 8865 Neutral and 5774 Negative tweets for the Demonetization dataset in summary, the approach proposed here appeared to be able to automate large Twitter corpora labeling for sentiment analysis, thus serving as a relevant alternative to time-consuming manual annotation, and provide grounds for future work on feature vector optimization and sentiment analysis on labeled datasets.

Gupta, Shi, Gimpel, and Sachan [20] presented a new method for the unsupervised induction of syntax and the interpretation of text representations. Note that the high-dimensional embeddings the authors are extracting from the likes of mBERT and E-BERT are fundamentally incompatible with typical clustering approaches like K-Means. To tackle this, they introduced SyntDEC, a deep clustering framework that simultaneously embeds the embeddings into a lower-cluster affinity space and conducts clustering. The research was conducted in two tasks on syntactic: part-of-speech induction (POSI), and constituency labelling (CoLab). Using such task-specific features (morphology for POSI, endpoint concatenation for CoLab), the approach successfully extracted syntactic information from the monolingual as well as the multilingual representations, unsupervised. On the 45-tag Penn Treebank WSJ dataset and the 12-tag multilingual universal treebanks, performance is competitive with or state-of-the-art, suggesting that pretrained language models have considerable syntactic knowledge built in. In addition, the method showcased its versatility by carrying out zero-shot syntax induction in the case

of resource-poor languages. In summary, our contribution presents a novel unsupervised probe for syntax built through deep clustering and extensive feature engineering to unveil the syntax learned by contextualized embeddings.

As presented above, recent studies of references from [10] to [16], the earlier efforts clustered English corpora with TF-IDF and standard algorithms. Earlier work in document clustering dealt with English corpora with natural TF–IDF representations and used mainstream document clustering algorithms like K-Means, HAC, and Affinity Propagation. Previous research in document clustering has focused on English corpora with standard TF–IDF representations and popular algorithms such as K-Means, HAC, and affinity propagation. Previous work obtained low purity for small data (36.7–52.5%) and does not scale, with very low Silhouette scores. Other works scaled the clustering to bigger corpora; however, only simple preprocessing and TF–IDF generated topics having low NMI, ARI, and Silhouette. Some dimensions of the design and mining constraints need to be reduced by dimensionality reduction methods such as hybrid optimization, but this leads to high computational cost, and is not suitable for large-scale data extractions. In comparison with previous work, our work of the POS-aware TF–IDF and K-Means clustering on the Kurdish Badini UOZBDN dataset leads to significantly outperforming baseline methods with the scores Purity 0.9714, NMI 0.9477, and ARI 0.9267 for higher cluster coherence, robustness, and computational efficiency. This demonstrates the advantage of integrating syntactic clues for better and more scalable document clustering, particularly for low-resource languages, as shown in Table I.

TABLE I. SUMMARY OF THE RELATED STUDIES ON DOCUMENT CLUSTERING

Ref. & Year	**Corpus**	**Approach**	**Techniques**	**Metrics (Purity ,NMI , Silhouette)**
[10] 2023	IMDB/Wiki (100) txt_Sentoken NLTK_Brown	Affinity Propagation for small datasets, K-Means for large datasets.	Affinity Propagation (AP), K-Means	Purity: 1, 0.624 Silhouette: 0.0475,0.0188
[11] 2020	Datasets: IMDB (50,000 reviews), Reuters (10,788 articles), and 20-Newsgroups (20,000 posts).	Document Clustering	K-means, Ward's Method	K-means, Ward's Method | Purity: 0.37–0.621 | NMI: 0.034–0.621 | Silhouette: 0.0006–0.0197
[12] 2020	IMDB, Wikipedia, 20-Newsgroups, txt-Sentoken	HAC, K-Means	Hierarchical Agglomerative Clustering (HAC) K-Means Clustering	Purity: 0.1059 , 0.6105 NMI: 0.0520 , 0.0350 |Silhouette: 0.0780 ,0.0139
[13] 2020	20-Newsgroup	Dimensionality reduction followed by clustering	K-means, SVD with K-means, NMF with K-means	NA
[14] 2021	English WordNet	Type 2 Intuitionistic Fuzzy Clustering and Seagull Optimization Algorithm (Type 2 IFCSOA)	FCM, FCM-PSO, FCM-GA, K-means	Purity: 0.65, 0.62, 0.61 NMI: NA Silhouette: NA
[15] 2023	Synthetic invoices & scanned pages (5 datasets)	Unsupervised document clustering using embeddings	k-Means, DBSCAN, HDBSCAN, BIRCH	Purity: N/A NMI: 0.853–0.8908 Silhouette: 0.4763
[16] 2025	Amazon Product Reviews	Unsupervised clustering	K-Means	Purity: 0.60–0.65 NMI: N/A Silhouette: 0.025–0.225

Several studies from [17] to [20] have explored the use of clustering techniques for unsupervised analysis of text and syntactic structures. In particular, clustering methods have been applied to text representations to induce Part-of-Speech (POS) tags. By leveraging morphological features, context-aware embeddings, and based dimensionality reduction, these models have demonstrated strong performance on both fine-grained and coarse-grained POS induction tasks across multiple languages. The unsupervised nature of these methods provides a less-biased alternative to supervised probing, enabling syntax induction even for resource-poor languages as indicated in Table II.

TABLE II. SUMMARY OF THE RELATED STUDIES ON DOCUMENT CLUSTERING PERFORMED ON POS-TAGGED DATASETS

Ref. & Year	**Corpus**	**Approach**	**Techniques**	**Result**
[17] 2009	Swedish text sets (DN: ~6,400 news articles; Occasional texts dataset: ~42,000 short texts)	Text clustering using K-Means	Text representations:: word forms, lemmas, PoS variants, stoplist, feature selection by word classes	DN NMI 0.44→0.52 (lemmatization) ; Occ NMI 0.10→0.25 (lemmatization + stoplist)
[18] 2024	English *Tafseer* translations of *Surah Al-Baqarah* (5 translators: TR1–TR5)	Clustering-based prioritization of translations	Preprocessing (tokenization, PoS, stemming, stop-words, normalization), TF-IDF/VSM, K-Means/AHC, SC/DBI/ET, PCA	K-Means consistent ranks (1,3), Agglomerative inconsistent, MCDM required for unified ranking

[19] 2021	Real-time Twitter datasets (#Demonetization, #Lockdown, #9pm9minutes)	Automatic labelling using improved K-means clustering	Feature engineering (POS, n-grams, lexicon), negation handling, TF-IDF/VSM, manual labels (Pos/Neg/Neutral)	Demonetization (4976/8865/5774), Lockdown (5189/8252/4924), 9pm9minutes (2405/2339/1614), improved silhouette & inertia, negation handled
[20] 2022	English WSJ / Universal Treebanks	Unsupervised POS induction & syntax probing	SyntDEC: mBERT/E-BERT, SAE, K-Means init, morphology features	POSI: WSJ 79.5%/73.9%, multilingual 75.7%, unsupervised syntactic learning

III. PROPOSED APPROACH

The proposed approach is implemented with the pipeline that starts with loading the Kurdish corpus and then progresses through four main stages: pre-processing, vectorization, dimensionality reduction, clustering, and evaluation, as follows:

A. Data collection

At first, the biggest challenge this work encountered when collecting data was the significant difficulty in finding standardized content in the Badini dialect, especially from online news or social media sources. Consequently, this work decided to utilize a corpus based on data from the UOZBDN corpus, which would be suitable for our study goals. Corpora were preprocessed and published in two ways: one annotated with POS tags and the other without any annotations. Of the POS tag, containing 51,761 POS-tagged tokens across 231 documents, named UOZBDN. The corpus was extracted from five thematic categories that include: Economy (ECO), Health (HEA), Politics (POLA), Society (SOC), and Sports (SPO). Each category was represented by a separate file containing multiple documents. The documents were POS tagged using a 38-tag Stanford POS tag set, including extensions and rules specific to the Badini dialect, as shown in Table III.

TABLE III. THE UOZBDN CORPUS SPECIFICATIONS.

No.	Topic	No. of Articles	No. of Sentences	No. of Tokens
1	Economy	46	189	10,458
2	Health	34	197	10,916
3	Political	37	145	10,338
4	Social	63	229	11,141
5	Sport	51	161	10,199
Total	**5**	**231**	**921**	**53,052**

B. Data Preparation

This experiment, loading the documents in the notebook (Google Colab), the function handles the files that include POS tagged documents, each file contains documents that start with a line, each word with its tag, ending with a period (.), and starts another document with a delimiter (#). This work separates each document into distinct categories (document0 for Economy, document1 for Health, document2 for Politics, document3 for Society, and document4 for Sports) to indicate the true label and, time, a numeric label corresponding to the document category is assigned from the filename and stored in a label list as well. Finally returns all cleaned and POS-filtered documents with their respective category labels. Along with the corpus that is tagged based on POS tagging, a corpus was created with POS tags for each word, and another file that contains the text as the documents are input into the process, No POS tags, separated by doc0, doc1, etc. The full content of each document was read line by line, were extracted. These words were all appended to each other, separated by a space to form a string to represent each document. Similar to the POS-tagged version, each document also receives a numeric category label, extracted from the file name, which is saved in another corresponding list of labels.

C. Part of Speech Tags Example

Ultimately, to compare it with the predicted label. The words from each document of the POS tag include 38 tags that contain nouns, verbs, adjectives, conjunctions, etc. From each document, select words into a list of documents that are separated by a space and store them in a string. At the same A part-of-speech tag is a label that indicates the grammatical role of a word in a sentence, such as a noun, verb, adjective, adverb, or pronoun. The following simple example from the Kurdish Badini dialect is included to illustrate how POS tagging contributes to linguistic analysis and sentence structure. In the Kurdish Badini sentence "ئەڤ ئەسمانە یێ شینە", each token is assigned a descriptive tag based on the UOZBND corpus annotation scheme, as the Table IV.

TABLE IV. POS TAGS EXAMPLE ACCORDING TO THE UOZBDN CORPUS (ئەڤ ئەسمانە یێ شینە).

Tag descriptive	Tags	Tokens
Demonstrative Adjective	JJD	ئەڤ
Noun	NN	ئەسمانە
Possessive Pronoun	PRPP	یێ
Descriptive Adjective	JJC	شینە
Sentence Break Punctuation	sent	.

The word "ئەڤ" is tagged as a demonstrative adjective (JJD), as it modifies and specifies the noun. The word "ئەسمانە" is labeled as a noun (NN), representing the main subject of the sentence. The word "یێ" is tagged as a possessive pronoun (PRPP), functioning as a linking element between the subject and its description. The word "شین" is identified as a descriptive adjective (JJC), providing a quality or attribute of the noun. Finally, the period is tagged as sentence break punctuation (sent). This tagging illustrates how the UOZBND corpus captures both grammatical function and descriptive roles of words in the Kurdish Badini dialect, supporting more precise linguistic analysis and feature extraction for natural language processing tasks.

D. POS Tag Filtering Strategies

POS-based feature selection methods were tested in this work to test their efficiency in Kurdish document clustering. The reason for doing this is that, in most cases, not all words have an equal contribution to identifying the contents of a document. Since pronouns, prepositions, conjunctions, etc., are primarily grammatical without much topical meaning, and because nouns,

verbs, and to some lesser extent adjectives (and adverbs) are far more informative. This work would simply divide the filtering of POS tags, focus on five cases: (ALL_TAGS, 22_TAGS, NVAA_TAGS, NV_TAGS, N_TAGS).

To systematically examine the effect of POS tagging on clustering Kurdish documents, five experimental cases were designed and evaluated through internal and external metrics. As an internal metric for assessing the cohesion and separation of clusters, the Silhouette Score was used, while for assessing the clustering quality concerning the ground truth, Purity, and NMI were used as external metrics.

- The first case, ALL_TAGS, comprised all the words within the 38 POS tags in the whole set.
- In the second case, 22_TAGS, only those syntactic categories were considered to be the
- most semantically meaningful when it comes to topic, for example, nouns, verbs, adjectives, adverbs, conjunctions, numerals, and interjections.
- Including specific tags: Use pos tags (NN, NNP, PRPD, JJC, JJD, JJI, JJT, JJF, VBT, VBI, RBT, RBP, RBC, RBD, RBF, INF, CC, PC, UH, NUM, CNUM, FW).
- In the third filter, NVAA_TAGS, only retained text categorized under noun, verb, adjective, and adverb tags to quantify the semantically rich categories that affect topic and context distinction. Including the words and their tag for the POS tags: Nouns (NN, NNP), Verbs (VBT, VBI, INF), Adjectives (JJ*), Adverbs (RB*).
- The fourth case, NV_TAGS, consisted of nouns and verbs only, creating a balance between topical ideas and their actions.
- Including Nouns (NN, NNP), Verbs (VBT, VBI, INF).
- Finally, N_TAGS, the fifth case, limited the representation to a set of nouns only, highlighting objects and subjects and points of concern (topics) as the basic markers of document semantics. Including Nouns (NN, NNP).

The current study thus systematically compared the impact of broader versus narrower POS-based filtering on clustering performance by testing five configurations and tried to reveal the most helpful linguistic features for Kurdish text clustering. By comparing these five configurations, the study systematically investigates how selective POS tagging, combined with preprocessing to remove delimiters, punctuation, and irrelevant tokens, can improve clustering performance and generate more coherent, topic-aligned clusters.

E. Document Clustering

According to the previous analysis of different clustering algorithms applied to investigate which is better for our model, it is concluded that K-Means achieves better performance than other clustering algorithms in POS tag enhanced document clustering for the Kurdish language corpus. In detail, K-Means gave the highest Purity (0.9714), NMI (0.9477), and Silhouette Score (0.2347). All of these metrics indicate overall high membership coherence, label suitability coherence, and cluster separation. On the contrary, the performance of the Fuzzy C-Means was much inferior to that, yielding a Purity of 0.4000, NMI of 0.4021, and a Silhouette Score of 0.1604, showing that the clusters were much weaker in structure and class consistency. Despite being a good alternative for clustering, which focuses on merging data points based on their distance, Agglomerative (Hierarchical) clustering had reasonably moderate results (Purity of 0.8286 and NMI of 0.7472, and 0.2347), as the K-means score nevertheless did not achieve K-Means performance. Overall, K-Means is the superior clustering method for the task, especially when augmented with POS tag information, and represents a robust baseline to build on for future improvements in low-resource languages' unsupervised clustering. So, K-Means was used for clustering documents with the centroid-based clustering algorithm that partitions the data into non-overlapping clusters in this research as a method to fit our data distribution and data structure like Table V.

TABLE V. PERFORMANCE COMPARISON OF CLUSTERING ALGORITHMS (K-MEANS, FUZZY C-MEANS, AND AGGLOMERATIVE) BASED ON PURITY, NMI, AND SILHOUETTE SCORE OF THE UOZBDN CORPUS

Clustering Algorithm	Purity	NMI	Silhouette Score
K-Means	0.9714	0.9477	0.2347
Fuzzy C-Means	0.4000	0.4021	0.1604
Agglomerative	0.8286	0.7472	0.2347

F. The Evaluation Metrics

In the final step, the quality of clustering results was evaluated using both internal and external metrics, consistency between predicted clusters and actual labels, and to help understand how effectively the proposed approach of enhancing the cluster with POS tags works on Kurdish documents. To evaluate these performances, external matrices (e.g., Purity, NMI) were used; the metrics indicate how well the labels conform to the clusters. Simultaneously, this work also employed the Silhouette Score as an internal evaluation metric to assess the levels of cohesion and separation within clusters. Well-separated and dense clusters have a higher silhouette score.

G. Diagram of All Steps in Methodology

The diagram of all steps in document summarization using POS tags is shown in Fig. 1.

1. Data Loading: Economy, Health, Politics, Society, and Sports
2. Selecting POS Tag Filtering
   - a. ALL_TAGS: All 38 POS tags
   - b. 22_TAGS: Selected 22 key syntactic/semantic tags
   - c. NVAA_TAGS: Nouns, Verbs, Adjectives, Adverbs
   - d. NV_TAGS: Nouns and Verbs only
   - e. N_TAGS: Nouns only
3. Data Splitting: Training (85%) and Testing (15%)
4. Feature Extraction (TF-IDF Vectorization): Documents into vectors
5. Dimensionality Reduction (SVD): To reduce vector space dimensionality

6. Clustering (K-Means Algorithm): K-Means (k = 5), FC-Means, HAC
7. Evaluation: External metrics and Internal metrics

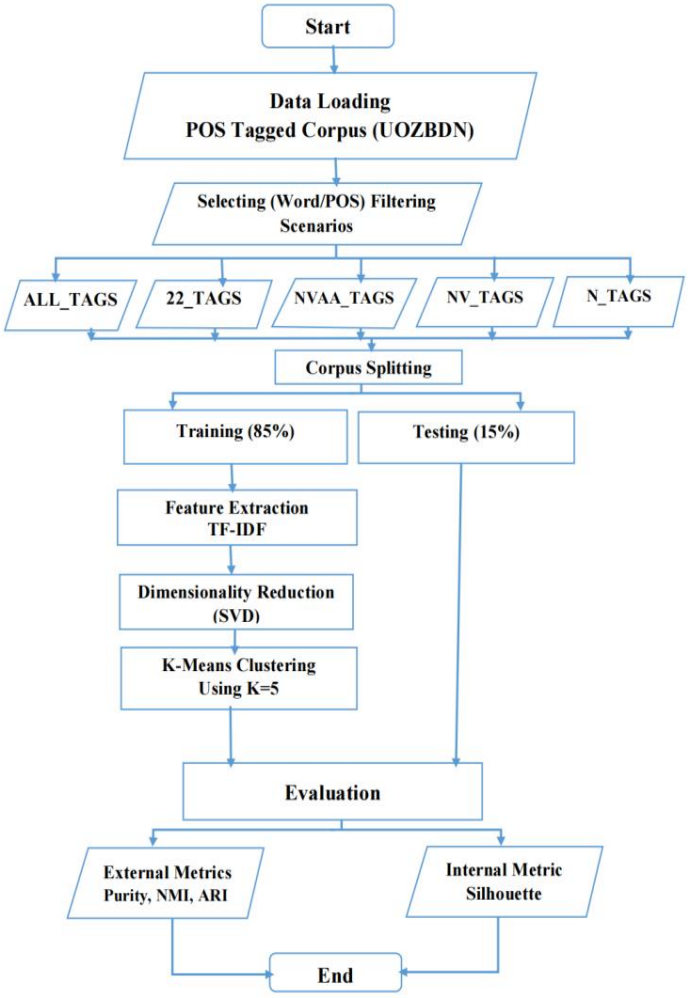

Fig. 1. Document Clustering Proposed Approach

IV. Experimental Results

The computational evaluation was used in five cases for evaluation. Clustering quality was evaluated using internal metrics and external metrics. Silhouette Score was used as the internal metric, while Purity and Normalized Mutual Information (NMI) were applied as external metrics to evaluate which is the better result. The five use cases of POS tags are as follows, as shown in Table VI.

TABLE VI. PURITY FOR VARIOUS POS TAG USAGE CASES.

Cases of Using POS Tags	Accuracy Result (Purity)
ALL_TAGS	0.8889
22_TAGS	**0.9714**
NVAA_TAGS	0.9143
NV_TAGS	0.7714
N_TAGS	0.8000

The results indicate that the 22_TAGS configuration achieved the highest clustering accuracy (Purity = 0.9714), outperforming the other POS-tag filtering strategies. This finding suggests that selecting an optimal subset of POS tags can significantly improve clustering performance, shown as a chart in the below Fig. 2.

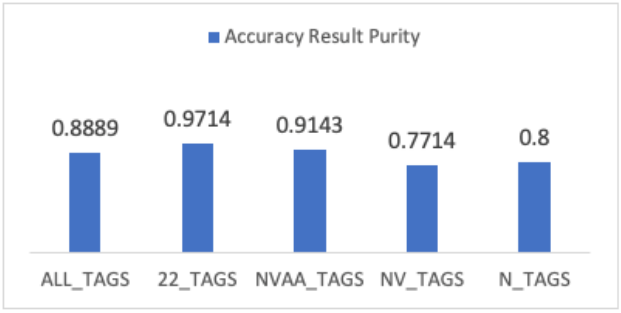

Fig. 2. Comparison of Purity Scores across five different POS-tag filtering strategies, demonstrating optimal cluster separation

The 22_TAGS configuration achieved the highest NMI score (0.9477), indicating a stronger agreement between the predicted clusters and the true document categories compared with the other POS-tag filtering strategies, as the Fig 3.

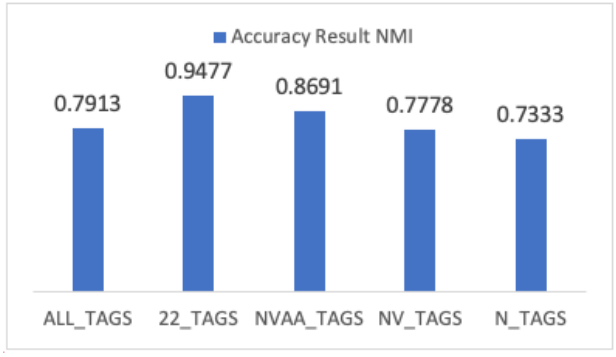

Fig. 3. Comparison of NMI Scores across five different POS-tag filtering strategies, demonstrating optimal cluster separation

TABLE VII. VARIOUS POS TAG USAGE CASES OF NMI

Cases of Using POS Tags	Accuracy Result NMI
ALL_TAGS	0.7913
22_TAGS	**0.9477**
NVAA_TAGS	0.8691
NV_TAGS	0.7778
N_TAGS	0.7333

The Silhouette scores for each possible configuration of POS tags are illustrated. The results indicate that the N_TAGS configuration achieved the highest Silhouette Score (0.2404), closely followed by the 22_TAGS configuration (0.2403), suggesting better cluster separation compared to the other POS-tag filtering strategies, as shown in Fig 4.

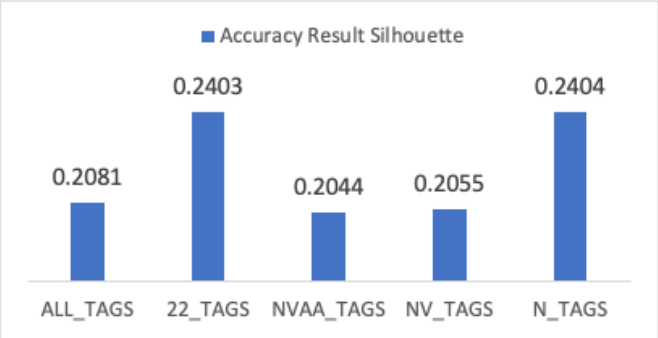

Fig. 4. Comparison of Silhouette Scores across five different POS-tag filtering strategies, demonstrating optimal cluster separation.

TABLE VIII. SILHOUETTE FOR VARIOUS POS TAG USAGE CASES.

Cases of Using POS Tags	Accurac of Silhouette
ALL_TAGS	0.2081
22_TAGS	0.2403
NVAA_TAGS	0.2044
NV_TAGS	0.2055
N_TAGS	0.2404

Note that, shown in Fig 5, comparative analysis between the 22_POS tags (22_TAGS) usage and (Not using POS), findings indicate that with POS tags, the clustering performance is improved substantially for all the evaluation metrics. Purity had the biggest percentage increase, going from 0.8857 to 0.9714, which means that with both algorithms, this work has a better homogeneity of clusters. Additionally, the NMI score increased from 0.8219 to 0.9477, coinciding with a better fit of predicted clusters with the ground truth. In addition to this, the Silhouette Score increased from 0.1725 to 0.2403, indicating the clusters are more compact/closer to each other. These results reinforce the compelling evidence through the selective integration of POS tags that is our configurations, namely 22_TAGS, further enhancing the performance of clustering as an unsupervised document cluster, as in Fig. 5.

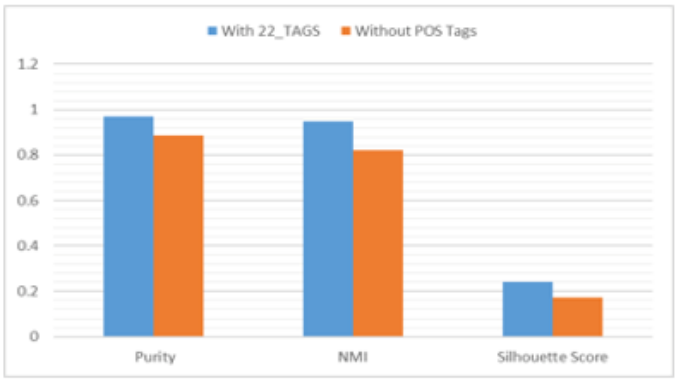

Fig. 5. Evaluation Scores (of Purity, NMI, and Silhouette Score) for both cases with 22_TAGS and without POS Tag.

TABLE IX. EVALUATION SCORES WITH 22_TAGS AND NO POS TAGS USING K-MEANS METHOD

Metric	With 22_TAGS	No POS Tags
Purity	0.9714	0.8857
NMI	**0.9477**	0.8219
ARI	**0.9267**	0.7630
Silhouette Score	**0.2403**	0.1725

V. DISCUSSION

Experimental results confirm that adding carefully selected grammatical information significantly improves the clustering of Kurdish documents. Using all words without POS filtering yields a purity of 0.8857, an NMI of 0.8219, and a Silhouette of 0.1725, demonstrating that lexical cues alone are not sufficient to fully separate the five topical categories. When including every tag (ALL_TAGS), frequent function words add additional noise and reduce the outlier scores to a purity of 0.8889 and an NMI of 0.7913, with a slight increase for Silhouette. In contrast, the (22_TAGS) subset—nouns, verbs, main adjectives, adverbs, numbers, foreign terms, and minimal coordination tags—raises performance to a purity of 0.9714, an NMI of 0.9477, and a Silhouette of 0.2403. Therefore, a balanced POS tag subset provides the optimal balance between information content and noise reduction.

The intermediate filters further demonstrate how linguistic richness affects corpus quality. Adding modifiers while deleting numerals and coordination cues (NVAA_TAGS) still outperforms the baseline (purity 0.9143, NMI 0.8691, Silhouette 0.2044), but it loses some discriminating power, suggesting that adjectives and adverbials alone do not fully discriminate between topics like economics and sports, where numerical data and foreign loanwords are important. Restricting features to nouns and verbs (NV_TAGS) or nouns only (N_TAGS) removes many thematic cues; Purity scores drop to 0.7714 and 0.8000, and mismatch scores to 0.7778 and 0.7333, with names alone yielding the highest (Silhouette) value of 0.2404, reflecting tight but mislabeled clusters.

Overall, the measures show that grammatical selectivity retaining content-bearing tags while excluding most function words provides the most pronounced subject separation for this obese, low-resource Kurdish group. Better performance of selective POS subsets stems from semantic information distributed across different grammatical categories. In Kurdish, the words that contain the most topical and event-related meaning are nouns, proper nouns, and verbs; therefore, they are highly discriminative for clustering tasks. In contrast, function words like conjunctions, particles, and pronouns are dispersed across all domains and thus blend the boundaries of clusters when used indiscriminately. The fact that adjectives and adverbs can be informative in other contexts does not mean that they define a topic; therefore, they do not help to clearly separate clusters. The impressive results obtained with the 22_TAGS subset imply that an appropriate trade-off is reached by preserving content words that are rich in semantics and removing high-frequency functional elements, thereby generating clearer topical differentiations in a morphologically

rich language like Kurdish. Future work should focus on a larger Badini corpus and developing NLP tools (POS taggers, tokenizers, lemmatizers) alongside standardized orthography to enable reproducible research. Additionally, leveraging advanced embeddings (BERT, FastText) with suitable clustering methods and detailed POS-specific error analysis can improve performance and help preserve and standardize the Badini dialect. In this study several limitations can be included, such as:

- The Badini dialect is a low-resource variety in the field of Natural Language Processing (NLP).
- There is a limited availability of corpora, especially annotated corpus for training and evaluation.
- Research methodologies specific to Badini are scarce, with very few existing studies.
- The absence of modern NLP tools and technique adaptations for the Badini dialect significantly restricts the development of reliable and high-performance NLP models for this linguistic variety.

VI. CONCLUSIONS

This research aimed to contribute to the field of Kurdish NLP by exploring the role of POS tagging in enhancing document Badini Kurdish language, the main challenges in this research is the absence of previous studies applying document clustering techniques to Kurdish corpora. Therefore, clustering documents in these under-resourced languages is highly challenging due to a lack of annotated corpora, morphological complexity, using UOZBDN corpora Incorporating POS tags in clustering, especially using just 231 documents across 5 different categories, compared different strategies for POS tag usage and found that using a selective approach with 22 key POS tags achieved the best performance across all evaluation metrics: Purity (0.9714), NMI (0.9477), and Silhouette Score (0.2403). These outcomes validate the importance of syntactic information in aligning with features related to each category and effectively differentiating documents semantically. The 22_TAGS configuration yielded a balanced, low-noise representation compared to POS information encoders. This research introduces a novel contribution by developing a POS tagged corpus for the Kurdish Badini dialect and employing it for document clustering. Incorporating POS information within the clustering framework leads to superior evaluation results, demonstrating that syntactic features can significantly enhance clustering performance for Kurdish language. No POS tag sets. The notable limitations of the study are the corpus size and dialect. Thus, future works include applying POS to larger Kurdish corpora in a multi dialect Kurdish corpus.

REFERENCES

[1] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," Artif. Intell. Rev., vol. 56, no. 7, pp. 6439–6475, July 2023.

[2] A. F. J. Al-Gburi, M. Z. A. Nazri, M. R. B. Yaakub, and Z. A. A. Alyasseri, "Multi-objective unsupervised feature selection and cluster based on symbiotic organism search," Algorithms, vol. 17, no. 8, Aug. 2024.

[3] A. Pegado-Bardayo, A. Lorenzo-Espejo, J. Muñuzuri, and A. Escudero-Santana, "A review of unsupervised k-value selection techniques in clustering algorithms," J. Ind. Eng. Manag., vol. 17, no. 3, p. 641, Aug. 2024.

[4] M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, "An improved k-means clustering algorithm towards an efficient data-driven modeling," Ann. Data Sci., Oct. 2022.

[5] A. Abas Abdullah, A. Mahmood Ahmed, T. Rashid, H. Veisi, Y. H. Rassul, B. Hassan, P. Fattah, S. A. Abdulhameed, and A. S. Shamsaldin, "Advanced clustering techniques for speech signal enhancement: a review and metanalysis of fuzzy c means, k means, and kernel fuzzy c means methods," CoRR, vol. abs/2409.19448, Sep. 2024.

[6] P. Safikhani and D. Broneske, "Enhancing AutoNLP with fine-tuned BERT models: an evaluation of text representation methods for AutoPyTorch," in Comput. Sci. Inf. Technol. (CSIT), vol. 13, pp. 23–38, Sept. 2023.

[7] A. M. Saeed, "An automated new approach in fast text classification: a case study for Kurdish text," Sci. J. Univ. Zakho, vol. 12, no. 3, pp. 330–336, June 2024.

[8] M. George and R. Murugesan, "Improving sentiment analysis of financial news headlines using hybrid word2vec–tfidf feature extraction technique," Procedia Comput. Sci., vol. 244, pp. 1–8, Jan. 2024.

[9] P. Morad, S. Ahmadi, and L. Gatti, "Part of speech tagging for northern kurdish," in Proc. Joint Workshop on Multiword Expressions and Universal Dependencies (MWE UD), Torino, Italy, pp. 70–80, May 2024.

[10] A. A. Mustafa and K. Jacksi, "Affinity propagation and k-means algorithm for document clustering based on semantic similarity," Sci. J. Univ. Zakho, vol. 11, no. 2, pp. 153–159, April 2023.

[11] N. M. Salih and K. Jacksi, "Semantic document clustering using k-means algorithm and ward's method," in Proc. 3rd Int. Conf. Adv. Sci. Eng. (ICOASE), pp. 1-6, Dec. 2021.

[12] K. Jacksi and N. Salih, "State-of-the-art document clustering algorithms based on semantic similarity," J. Informatika, vol. 14, no. 2, p. 58, 2020.

[13] R. Kumbhar, S. Mhamane, H. Pati, and S. Patil, "Text document clustering using k-means algorithm with dimension reduction techniques," in Proc. IEEE Int. Conf. Comput. Electr. Commun. Eng. (ICCES), pp. 1164-1168, June 2019.

[14] P. Perumal and B. Mathivanan, "Type2 IFC with SOA for topic detection and document clustering analysis," Research Square, 2021.

[15] R. Saha, "Influence of various text embeddings on clustering performance in nlp," arXiv preprint arXiv:2305.03144, 2023.

[16] P. R. Sampaio and H. Maxcici, "Unsupervised document and template clustering using multimodal embeddings," arXiv preprint arXiv:2506.12116, June 2025.

[17] M. Rosell, "Part of speech tagging for text clustering in swedish,", In proceedings of the 17th Nordic Conference of Computational Linguistics, pp. 150-157, May 2009.

[18] M. A. Ahmed, S. M. Nafl, H. Baharin, and P. N. E. Nohuddin, "Prioritise five tafseer translators using clustering technique for surah al-baqarah," Al-Iraqia J. Sci. Eng. Res., vol. 3, no. 1, pp. 75–86, March 2024.

[19] I. Gupta and N. Joshi, "Real-time twitter corpus labelling using automatic clustering approach," Int. J. Comput. Digit. Syst., vol. 10, no. 1, pp. 519–532, April 2021.

[20] V. Gupta, H. Shi, K. Gimpel, and M. Sachan, "Deep clustering of text representations for supervision-free probing of syntax," In Proceedings of the AAAI Conference on AI, Vol. 36, No. 10, pp.10720-10728, Jun. 2022