applied sciences

MDPI

Article

The Role of Transliterated Words in Linking Bilingual News Articles in an Archive

Muzammil Khan [1,*], Sarwar Shah Khan [1], Yasser Alharbi [2], Ali Alferaidi [2], Talal Saad Alharbi [2] and Kusum Yadav [2]

[1] Department of Computer and Software Technology, University of Swat, Mingora 19130, Pakistan; sskhan0092@gmail.com

[2] College of Computer Science and Engineering, University of Ha'il, Ha'il 55473, Saudi Arabia; y.alharbi@uoh.edu.sa (Y.A.); a.alfredi@uoh.edu.sa (A.A.); t.alharbi@uoh.edu.sa (T.S.A.); y.kusum@uoh.edu.sa (K.Y.)

* Corrsependence: muzammilkhan86@gmail.com

Abstract: Retrieving a specific digital information object from a multi-lingual huge and evolving news archives is challenging and complicated against a user query. The processing becomes more difficult to understand and analyze when low-resourced and morphologically complex languages like Urdu and Arabic scripts are included in the archive. Computing similarity against a query and among news articles in huge and evolving collections may be inaccurate and time-consuming at run time. This paper introduces a Similarity Measure based on Transliteration Words (SMTW) from the English language in the Urdu scripts for linking news articles extracted from multiple online sources during the preservation process. The SMTW link Urdu-to-English news articles using an upgraded Urdu-to-English lexicon, including transliteration words. The SMTW was exhaustively evaluated to assess the effectiveness using different size datasets and the results were compared with the Common Ratio Measure for Dual Language (CRMDL). The experimental results show that the SMTW was more effective than the CRMDL for linking Urdu-to-English news articles. The precision improved from 50% to 60%, recall improved from 67% to 82%, and the impact of common terms also improved.

Keywords: transliterated words; news archiving; news linking; dual lingual archive; digital libraries; similarity measure

Citation: Khan, M.; Khan, S.S.; Alharbi, Y.; Alferaidi, A.; Alharbi, T.S.; Yadav, K. The Role of Transliterated Words in Linking Bilingual News Articles in an Archive. *Appl. Sci.* **2023**, *13*, 4435. https://doi.org/10.3390/app13074435

Academic Editors: Adegboyega Ojo and Rizun Nina

Received: 3 February 2023
Revised: 27 March 2023
Accepted: 28 March 2023
Published: 31 March 2023

1. Introduction

A plethora of digital information is available from many sources, and the WWW (World Wide Web) is growing rapidly and is an essential and fragile source. According to a study, Google and Bing indexed about 5.47 billion web pages [1]. Search engines index and store approximately one hundred million gigabytes of digital information, and hundreds of gigabytes are added every day [2,3].

The web's fragile nature prompts them to disappear digital information quickly. Most of the digital information disappears , as much as eighty percent (80%) of webpages become unavailable after one year, and thirteen percent (13%) of references to scholarly articles and web links appear broken over a period of 27 months [3,4]. According to a Google survey, the people using Google search engine expect to load a webpage within two seconds, and they abandon it if it takes more than three seconds. The question remains as to how it will be if the webpage is not accessible at all [5]. Thus, information fragility causes this valuable information to vanish and become unavailable.

The worst-case may be the inaccessibility or loss of digital objects from online sources providing this information. Numerous different resources provide a variety of information to users. Digital information must be protected from being lost and preserved in a centralized or local digital collection [1]. A digital collection with a considerable volume is challenging to utilize and manage, whether online or offline, such as digital archives or any digital library.

A massive collection of online digital information for web users is available, including news articles, research articles, hotels, restaurants, blogs, movies, and opinions on various products in the form of forms or books, etc. However, the information provided by the news is one of the important types covering different aspects of life and an established source of knowing history. News is instantly generated activity published online, but the lifespan is very short. Hence, it is required to preserve this digital news for use in the future and ensure that the news remains accessible, usable, and available, as long as they are conceived as important [6].

There are many approaches introduced that preserve digital information, such as the preservation of research data [7] and the model migration approach that preserves databases [8]. However, the preservation of news is complicated and has many challenges because it is not easy to access related news articles from multi-source and multi-lingual news archives such as a Digital News Stories Archive [9]. The metadata helps to organize digital news objects in the digital archives or libraries and helps to locate, retrieve, manage, structure, and preserve the digital objects [10]. Linking mechanisms and metadata are highly important to ensure the dissemination of archived news articles extracted from multiple sources in multiple languages during the preservation process. Artificial intelligence (AI) has a significant impact on the accessibility of news or other digital information for a huge collection of multilingual archives using advanced natural language processing tasks. For example, to provide a personalized recommendation based on user interaction and browsing history of news reading behavior, machine learning technique-based language models can help to predict accurate searches in a multilingual environment for multilingual retrieval and query manipulation. Similarly, translation tools and techniques can help to manage digital content during the information dissemination process, which encompasses a number of challenges [11].

The use of English transliteration words is common in most low-resource languages such as Urdu and may have a great impact on linking digital content for dissemination purposes in the future. The main goal of this paper is to introduce a linking mechanism based on the use of English transliterated words in Urdu news articles, and it examines the impact of transliteration words in Urdu news articles to ensure the accessibility of news articles that are extracted and archived from multiple sources during the preservation. The linking algorithm is presented in detail for linking dual-language news articles. The proposed algorithm, i.e., SMTW, is evaluated using a hybrid evaluation method, such as evaluation of both user's centric and system-centric evaluation approaches, and the results are compared against the Common Ratio Measure for Dual Languages (CRMDL) to clearly formulate the impact of English transliteration words in Urdu scripts. The Digital News Stories Preservation (DNSP) framework is enriched with different linking mechanisms to ensure accessibility in the future.

The rest of the paper is organized as follows: Section 2 and its Sections 2.1 and 2.2 give the background of the DNSP framework, the DNSA, the contributions made to the framework, and the need for linking mechanisms. Section 3 discusses the proposed transliteration-based similarity measure for linking, provides a brief about transliteration words and the role of transliteration words in Urdu scripts, and gives comprehensive details about the dataset used for evaluation. Section 4 presents the results and comparison of the proposed algorithm with the CRMDL. The last Section 5 summarizes the paper's findings.

2. Background

The Digital News Stories Preservation (DNSP) framework was initiated to preserve digital news articles published online in the English language from different platforms that were then enhanced for multiple languages, i.e., Urdu, Arabic, and English [12]. The DNSP framework uses content-based techniques to preserve and create a multi-lingual news archive, i.e., the Digital News Stories Archive (DNSA) [13]. The archive is enabled to preserve news articles published online in two low-resource languages, i.e., Urdu and Arabic, and one high-resource language, i.e., English, from eighteen multiple news sources. The Digital News Story Extractor (DNSE) is an important component of the DNSP Framework that facilitates the extraction of news articles from online news publishing platforms, supports format migration, and normalizes news articles during preservation to DNSA.

2.1. Digital News Stories Archive (DNSA)

In this section, we are briefly introducing the Digital News Stories Archive (DNSA). The principal idea of the Digital News Stories Preservation (DNSP) framework is presented at the International Conference on Asian Digital Libraries 2015 (ICADL-2015) [12]. The following are the major contributions to the framework:

- A generic systematic approach was proposed as a web preservation model, i.e., a step-wise model for web preservation projects after analyzing 120 news archives worldwide [14,15].
- A multiple source web archive for online news articles, Digital News Stories Archives (DNSA), was created to preserve news articles from multiple sources [1].
- A tool “Digital News Stories Extractor (DNSE)” was developed to extract news articles from multiple sources to create the DNSA [13].
- Content-based techniques were introduced for linking news articles during the preservation process in the DNSA. These text processing techniques are based on text features, such as common ratio, terms frequency [16], named entities [17], term position, information credibility, headline terms, similar terms distance, etc. [1].
- The news recommendation techniques were studied comprehensively for similarity measures. The study helped identify various dimensions and enhanced the DNSP framework, and a few were identified for future research in the framework [18].
- The Common Ratio Measure for Stories (CRMS) technique was modified for linking English news articles during preservation and limited to news headings to reduce extra computation for the terms appearing in the news body [16].
- The CRMS technique was modified for linking dual languages, i.e., linking Urdu-language news articles with English-language news articles during preservation in the DNSA [19].
- A heading-based linking mechanism was introduced for the archived news articles during the preservation process in the framework [20].
- Recently, the framework has been enriched with news articles from the Arabic language. The challenges were identified for including low-resource languages, such as Urdu and Arabic languages, and a set of metadata was introduced to best serve the DNSP framework, which was adapted for multi-lingual news archives.

The digital news stories archive (DNSA) was created locally from multiple sources that preserve news articles published in English, Urdu, and Arabic, due to a lack of funds and support from institutes and funding bodies.

A news archive without efficient retrieval mechanisms will just be a collection of digital news objects, rather than a helpful information repository. Implementing an efficient search requires using indexing approaches, metadata, and linking mechanisms so that they help news readers retrieve relevant articles easily and effectively.

2.2. Linking Digital News Stories in DNSA

An immense collection of digital information for use by web users is online available, including news articles, research articles, hotels, restaurants, blogs, movies, and opinions on various products in the form of books, etc. Recommender systems help web users focus on the information they need that is provided in manageable units. Generally, the techniques used by the recommendation system is divided into the Collaborative Filtering approach, which is based on similar users having the same demographics or similar interest, and the Content-based approach, which is based on the features of the items [18,21,22].

The extraction trial shows that the extraction and preservation of available news articles can be huge, and recommendation systems can help recommend relevant news based on predefined criteria to filter news for the news readers. The collaborative or content-based approach can be adopted for linking news. The collaborative filtering technique faces several challenges, as it depends on the similarity in demographics and opinions of the users [23,24], and the dynamic nature of users makes it more complicated. In an online news environment, the users normally preferred to find recent news, which makes it hard to trace web users' preferences that lead to an accurate model based on the contents they previously read [25–27]. User interest changes over time, depending on news articles of the popular current events themselves [28]. Generally, during news reading, the users are not willing to recommend news during news searching and browsing [29]. Content-based approaches recommend new objects to the user based on the features of the object previously selected or the computed similarity value between the descriptions or meta-elements [30]. Content-based approaches can run through their problems, such as determining the similarity between news articles that represent different topics and the way the user's choice effect by some potentially hidden factors.

All these studies are focusing on the currently evolved news and compute run time similarity, which are mostly based on user queries. In our earlier study, different aspects related to recommendation systems and techniques that were mostly used in an environment of online news were discussed. For example, they included news sources conceived for experimental trials, datasets used, recommendation approaches, efficiency estimation, evaluation techniques, etc. [18].

3. Similarity Measure Based on Transliteration Words

3.1. Transliteration

"Transliteration is a process of using the text of one script in another script or the process of converting text from one language to another". Transliteration replaces words from a source language with the target language's spelling equivalents or approximate phonetics. In linguistics, the process through which a word or set of words of a language is adapted for use in another language's script is referred to as borrowing, and the word(s) are also known as loanwords [31]. Transliteration utilizing a phrase or word in a language with a distinct writing system [32] becomes more difficult if a language has a distinct sound and writing scripts [33].

Transliteration is not a translation in linguistics. In language translation, the written and spoken sense of the text or words in the target language is transferred from a source language. In contrast, in transliteration, the meaning of the words or text does not change or render, but only the source characters or letters change into a corresponding target language.

English Transliteration in Urdu Scripts

Most of the spoken languages acquire several words from other languages using different character sets. Similarly, native speakers of Urdu frequently use several words from other languages, especially from the English language. The English-based origin words are used with different characters and identical pronunciations, despite having alternative words in the Urdu language. As a considerable proportion of English transliteration words are used in Urdu, the effect of these words in Urdu news articles must be estimated for the link, especially for calculating similarity among news articles in the DNSA. Table 1 shows examples of transliteration words from English in the Urdu scripts.

Here are two examples of Urdu language sentences with underlined transliteration words: بجٹ تنخواہ و پنشن میں 10 فیصد اضافے کی تجویز has "budget" and "pension" as transliteration words, and یونس خان کی ‘‘آل ٹائم’’ ٹیسٹ ٹیم کے کپتان عمران خان has "all-time", "test", and "team" as English transliteration words used in Urdu language scripts.

Table 1. Example English Transliteration Words in Urdu Script.

English Word	English Transliteration Words	Urdu Word	Phonetic Transcript	Roman Urdu
New	نیو	نیا	nəja:	Naya
English	انگلش	انگریزی	əŋgrezi	Angrezi
Energy	انرجی	طاقت	t̪əˈqɑ:t̪	Taqat
School	سکول	مدرسہ	məˈdrɑ:sə	Madrasa

A sample of six hundred (600) Urdu news articles collected from different sources was analyzed to specify the use of English transliteration words in Urdu news articles using the DNSE. The stopwords were removed from the news articles during preprocessing, and the corpus contained a total of 117,393 tokens. The estimation was analyzed against a collection of 2705 English transliteration words. Table 2 summarizes the percentages of total tokens, Urdu words, English words, and Unique tokens in the corpus. Figure 1 shows that 9.5% are English transliteration words, 19.5% are other words (for example, symbols, digits, etc.), and 71% of the words in the Urdu news articles are Urdu origin words in the sample corpus.

Table 2. Tokens Distribution in 600 News Articles Corpus.

Token	Count	Percentage
Total Tokens	117,393	100%
Unique Tokens	10,914	9.2%
Total Urdu Words	101,147	86.1%
Unique Urdu Words	7770	6.6%
Total English Words	9962	8.4%
Unique English Words	1038	0.9%

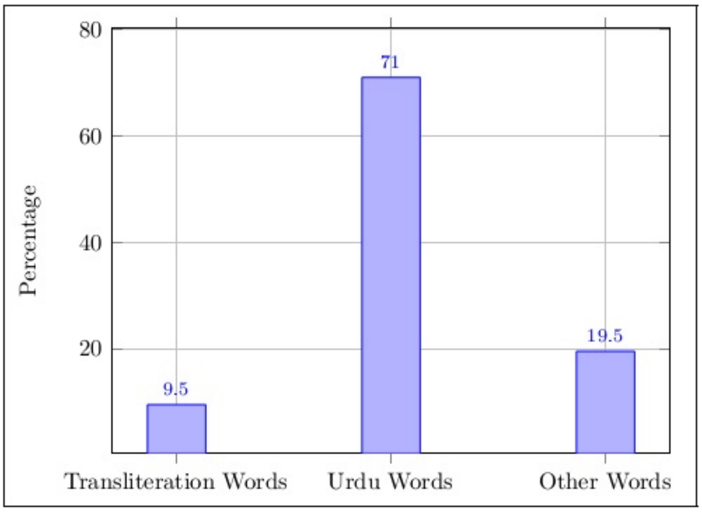

Figure 1. Transliteration words ratio to Urdu words in sample corpus.

3.2. Role of Transliteration Words

Transliteration words play a significant role in natural language processing tasks, depending upon the number of transliteration words used in that language. Almost all informal languages comprehend several transliteration words. In Urdu, a large collection of English transliteration words are frequently used both in spoken and written scripts by native speakers with the same characters and pronunciations, despite alternative words in the Urdu language being available.

A sample of six hundred (600) Urdu news articles from different sources were analyzed to specify the use of English transliteration words in the Urdu news articles' writings. A large portion of transliteration words were used in the Urdu scripts, which can help to link Urdu news articles with English news articles in the DNSA [34]. We introduced the following Algorithm 1 to show the effects of transliteration words on linking bilingual news articles.

Algorithm 1: SMTW Algorithm Pseudo-Code

```
Input: New News Article (NNA) and Archived News Articles (ANA) $\in$ DNSA
Output: Similarity Score of NNA with ANA
// NNA is an Urdu news and ANA are English news articles
1  NNA preprocessing // Filtering non-news contents and extracting the news article
   from the news webpage during extraction
2  Tokenize NNA
3  T = $t_1, t_2, t_3, ..., t_n$
4  Removing stopwords (if any)
5  Computing term frequency (TF) of each term in NNA
   // Find the English meaning of each Urdu word in Urdu-to-English Dictionary
   (U2E-Lexicon) Or in Transliteration words List (TWL)
6  for $w_i \in Map(NNA)$ do
       // If word exists in U2E-Lexicon
7      if $w_i \in$ U2E-Lexicon OR $w_i \in$ TWL then
           // For a single Urdu word, multiple meanings should identify; for example,
           for brtAw many words can be used in a similar English news article is,
           Behave, Behaviour, Conduct, Treat, etc.
8          Find TF of each term t from T
9          Populate Map(NNA) // Populate NNA Map for Word w and its frequency
10         Map(NNA) = $(tf_1, w_1),(tf_2, w_2),(tf_3, w_3),...,(tf_n, w_n)$
11     else
12         NewWord($w_i$) // Add new Urdu word to array NewWord
13 for ANA $\in$ DNSA do
       // Repeat steps 1 to 4, to Compute the term frequency (TF) for each term in ANA
14     for $w_i \in Map(ANA)$ do
15         Find TF of each term t from T
16         Populate Mam(ANA) // using ANA Map for Word w and its frequency
17         Map(ANA) = $(tf_1, w_1),(tf_2, w_2),(tf_3, w_3),...,(tf_n, w_m)$
18     Map(NNA) = $(tf_1, w_1),(tf_2, w_2),(tf_3, w_3),...,(tf_n, w_n)$
19     Map(ANA) = $(tf_1, w_1),(tf_2, w_2),(tf_3, w_3),...,(tf_n, w_m)$
       // CT - Common Terms, UT - Uncommon Terms, and TT - Total Terms
20     CT = $(tf_1, tf_2)w_1,(tf_1, tf_2)w_2,(tf_1, tf_2)w_3,...,(tf_1, tf_2)w_n$
21     CT = $\sum_{i=1}^{n} ((tf_1, tf_2)w_i)$
       // Where $W_1$ is the 1st common term or word, $W_2$ is the 2nd common term or word
       in both the selected news articles and so on, $W_i$ is the common term or word in
       both the news articles, $tf_1$ term frequency of word $W$ in Urdu news, $tf_2$ is term
       frequency of word $W$ in English news article and $n$ is the total number of
       common terms in both the news.
22     UT = $(tf_1, tf_2)w_1,(tf_1, tf_2)w_2,(tf_1, tf_2)w_3,...,(tf_1, tf_2)w_m$
23     UT = $\sum_{i=1}^{m} ((tf_1, tf_2)w_i)$
24     TT = UT + CT
25     SMTW = CT/TT or CT/UT
```

The proposed "similarity measure based on transliteration words" approach, i.e., SMTW, for computing similarity within news articles was analyzed using different datasets, as discussed in Section 3.3.

3.3. Datasets

Due to the continuous extraction of news articles from multiple sources, the DNSA can grow very quickly for both high- and low-resource languages. Approximately, four hundred (400) Urdu news from five (5) sources, one hundred and eighty (180) news from Arabic from three (3) sources, and seven hundred (700) English news articles from ten (10) online sources were extracted by the DNSE on a daily basis.

For evaluation, the heading or title of the news articles was read for the dataset selection from currently hot topics from the general pool. A brief overview of the datasets used for evaluating the proposed similarity measures is presented in Table 3.

Table 3. Datasets Overview Bilingual News Articles.

			News Articles			Similarity Observed		
No.	**News Articles/Set**	**Sets**	**Urdu Articles**	**English Articles**	**Sources**	**During Selection**	**Proposed Measures**	**Results Observed**
1	4	3	1	3	3	Yes	Yes	Yes
2	10	2	5	5	5	Yes	Yes	Yes
3	20	1	10	10	5	Yes	Yes	Yes
4	282 (One Day)	2	152	130	4	No	Yes	Yes

The selection of news articles for the dataset and the selection criteria were informed and closely analyzed for the proposed linking mechanisms introduced in [16,17,19,20].

The datasets used for the evaluation of the proposed similarity measure are briefly discussed below:

- **Four news article sets—**each set contains one Urdu and three English news articles in which one Urdu news article is similar to one English news article, and the two news articles are selected differently from other sources. The news is keenly analyzed, and the similarity score is computed for the SMTW technique during the implementation. Tokenization, identification, and extraction of the transliteration words and preprocessing of Urdu news articles are observed during the implementation of the proposed algorithm.
- **Ten news articles set—**each set contains five (5) English news articles that are similar to five (5) Urdu news articles and is used to observe the problems encountered, such as matching and missing terms during matching transliterated words, the effects of capitalization of words, etc., as well as improving the structure of the dictionary, including all possible transliteration words. Each set contains five English and five Urdu news articles.
- **Twenty news articles set—**contains ten (10) English news articles that are similar to ten (10) Urdu news articles and is used to compare the outcome of the proposed similarity technique. The news article sets are used to improve the structure and contents of the Urdu-to-English lexicon for transliterated words and related structure issues of Urdu scripts.

  Similar articles are selected in both languages by reading the heading or title of the news articles for the twenty news dataset selection from currently hot topics from the general pool. Similar news articles are named Ur1, Ur2... Ur<n> and Eng1, Eng2, . . . Eng<n>. It contains five national and international news articles, five sports news articles, and one sport plus national news article, as presented in Table 4.

- **A set of 282 news articles** is used to observe the overall effects of the proposed similarity measure. The news is extracted from two online television broadcasters, i.e., Geo and Samaa news, in both the English and Urdu language. The collection contains one hundred and fifty-two (152) Urdu news articles and one hundred and thirty (130) English news articles from the general pool. The set of news articles used for empirical evaluation is summarized in Table 5.

Table 4. Overview: 20 news article dataset in dual languages.

Type of News	News Articles	News Articles	About
Sports News	3	1, 6, 10	PSL, Cricket
	2	7, 9	WI tour, Teams announcement
	1	5	ICC president resign
General News	3	2, 6, 8	COAS, Army
	1	3	Trump travel ban
	1	4	MQM leader

Table 5. News articles to be analyzed for similarity.

Urdu Article	بجٹ 2017ـ18؛ سرکاری ملازمین کو خوش کردیا گیا
Description	Having no exact match, much similar news, general news, and of average length
Stats	6 relevant news out of 55 and no exact match
Urdu Article	رمضان المبارک کا چاند نظر آگیا، پہلا روزہ کل ہوگا
Description	Having no exact match, much similar news, general news, and of short length
Stats	9 relevant news out of 55 and no exact match
Urdu Article	بجٹ تنخواہ و پنشن میں 10 فیصد اضافے کی تجویز
Description	Having one exact match, much similar news, general news, and of average length
Stats	8 relevant news out of 74 and one exact match
Urdu Article	یونس خان کی ’’آل ٹائم‘‘ ٹیسٹ ٹیم کے کپتان عمران خان
Description	Having one exact match, much similar news, sports news, and of average length
Stats	7 relevant news out of 74 and one exact match

4. SMTW Evaluation

It is observed that native speakers of the Urdu language use many English transliteration words frequently in both written scripts and in the spoken language. The "Common Ratio Measure for Dual Languages (CRMDL)" is a team-based approach, which was modified to a "Similarity Measure based on Transliteration Words (SMTW)" to improve dual lingual linking accuracy among news articles in the DNSA. The proposed technique was analyzed and compared with the CRMDL empirically via datasets presented in Table 3.

4.1. Results

The similarity was computed by implementing the SMTW and was analyzed vigorously to assess the worth of the proposed approach. The common ratio CT/TT shows reliable and promising results as compared to the UT/TT common ratio and, hence, was included for evaluation. The results of twenty news article sets highlighted for the SMTW are presented in Table 6 below.

The proposed similarity measure of the SMTW shows encouraging results for all Urdu news articles by comparing relevant English news articles. The results of each Urdu news article were ranked and comprehensively compared to observe the effectiveness of the English transliteration words used in the Urdu news articles.

Table 6. Computed Similarity for 20 News Articles using SMTW

UrduNews	EngNews	CRMDL	CT
ur1	eng1	0.254	18
	eng6	0.118	13
	eng10	0.113	13
	eng7	0.065	20
	eng4	0.035	6
ur2	eng2	0.191	37
	eng6	0.054	11
	eng8	0.049	12
	eng5	0.044	9
	eng3	0.038	9
ur3	eng3	0.111	25
	eng4	0.048	12
ur4	eng4	0.105	41
	eng7	0.028	15
ur5	eng5	0.14	37
	eng1	0.072	16
ur6	eng6	0.255	41
	eng2	0.098	15
	eng8	0.078	16
	eng1	0.066	8
	eng7	0.064	23
ur7	eng7	0.155	98
	eng10	0.126	55
	eng6	0.121	52
ur8	eng8	0.155	38
	eng6	0.094	19
	eng2	0.062	12
	eng4	0.05	13
	eng3	0.034	8
ur9	eng9	0.165	38
	eng7	0.108	49
	eng6	0.059	15
ur10	eng10	0.192	42
	eng6	0.112	24
	eng1	0.097	17
	eng7	0.08	33
	eng9	0.042	8

The results presented in Table 7 showed the effectiveness of the SMTW for linking Urdu-to-English news articles for individual broadcasting sources. The first column "Rank" in the table represents the similarity rank of each similar news article in the dataset, the second column represents news labels that use acronyms to use limited space efficiently, the third column presents the SMTW value, and the fourth column represents the common terms among Urdu and English news articles.

Precision and Recall

The precision and recall evaluation matrices were computed to analyze the accuracy of the SMTW measure. The experimental results were obtained from a one-day dataset which contained two hundred and eighty two (282) news articles extracted from four news sources. The relevant news and features of the news articles were specified, such as the length of news, much similar news, exact match news, and the number of relevant news articles, as shown in Table 5. The computed precision and recall experimental results are shown in Table 8.

A "similarity measure based on transliteration words (SMTW)" seems feasible for calculating the content-based similarity for linking Urdu-to-English news articles during the preservation process. The SMTW is better for lengthy news articles than for short news and more feasible for sports news. The digital news stories archive preserves linked and formatted news articles to ensure that the related news articles were accessible in the future from an enormous corpus of news articles extracted from multiple sources using the SMTW measure.

Table 7. Computed similarity for one day news articles using SMTW.

UrNews	بجٹ 2017ـ18؛ سرکاری ملازمین کو خوش کر دیا گیا		
Rank	**Relevant English News**	**SMTW**	**CT**
1	Eng1	0.25	75
2	Eng2	0.18	31
3	Eng3	0.17	31
4	Eng4	0.17	36
7	Eng6	0.12	20
6	Eng5	0.12	34
UrNews	رمضان المبارک کا چاند نظر آ گیا، پہلا روزہ کل ہو گا		
Rank	**Relevant English News**	**SMTW**	**CT**
1	Eng1	0.12	18
2	Eng2	0.09	14
4	Eng3	0.04	6
9	Eng4	0.04	4
12	Eng6	0.03	6
13	Eng5	0.03	7
17	Eng7	0.02	6
19	Eng8	0.02	11
26	Eng9	0.02	2
UrNews	بجٹ تنخواہ و پنشن میں 10 فیصد اضافے کی تجویز		
Rank	**Relevant English News**	**SMTW**	**CT**
1	Eng1	0.26	121
2	Eng2	0.22	115
3	Eng3	0.19	219
4	Eng5	0.18	97
5	Eng8	0.17	73
6	Eng4	0.17	106
7	Eng7	0.17	86
8	Eng6	0.16	83
UrNews	یونس خان کی ''آل ٹائم'' ٹیسٹ ٹیم کے کپتان عمران خان		
Rank	**Relevant English News**	**SMTW**	**CT**
1	Eng2	0.18	176
2	Eng1	0.17	122
3	Eng5	0.14	103
4	Eng3	0.10	81
7	Eng7	0.09	71
9	Eng4	0.09	69
11	Eng6	0.07	50

Table 8. Precision and recall for SMTW.

Urdu News	Precision	Recall
بجٹ 2017-18؛ سرکاری ملازمین کو خوش کر دیا گیا (Budget 2017–18: Government employees were made happy)	60%	100%
رمضان المبارک کا چاند نظر آ گیا، پہلا روزہ کل ہو گا (The Ramadan moon sighted, the first fast will be tomorrow)	40%	44%
بجٹ تنخواہ و پنشن میں 10 فیصد اضافے کی تجویز (Budget, 10% raise in salaries and pension)	80%	100%
یونس خان کی ''آل ٹائم'' ٹیسٹ ٹیم کے کپتان عمران خان (Yonus Khan's all-time test captain is Imran Khan)	60%	86%
Average	**60%**	**82%**

4.2. Common Ratio Measure and Transliteration Words Measure Comparison

The content-based techniques "CRMDL" and "SMTW" performed well for linking a low-resource language, i.e., Urdu, and a high-resource language, i.e., English. The SMTW was compared against the CRMDL and keenly analyzed, and the improvement imparted by the SMTW is highlighted in this section. The comparison is made for three evaluation parameters, which are:

1. Result Improvement
   The results of both the CRMDL and SMTW were compared, and the improved results of the SMTW were highlighted and ranked. Improvement means that the result includes all the relevant news in the top-five news or the rank of the relevant is improved, i.e., the most relevant news brought to the top of the top-five news articles. In contrast, "Dropped" means a similar news article in the top five is fallen, and "None" is used for the same results in both techniques or for no effect by the new technique.
2. Transliteration Words Impact
   The use of English transliterated words is frequent in Urdu scripts and will surely have an impact on the count of common terms. The impact of transliteration words on the results was analyzed and showed the effects of linking Urdu and English news articles.
3. Result Accuracy (precision and recall)
   The results' accuracy needs to be compared in terms of precision and recall for both dual-lingual news articles and to assess the overall feasibility of the proposed similarity measure.

Table 9 shows the dominance and better performance of the SMTW over the CRMDL for linking Urdu news articles with relevant English news articles during the presentation and development of the DNSA. The transliterated words played an important role in computing the similarity value among relevant news in multi-lingual archived news articles. The similarity improved by 22%, i.e., 5 out of 23, in which ranking improved by 13% and results improved by 09% for relevant news. The result remained unchanged by 74%, and the computed similarity dropped by 04% for Urdu news ur6 only.

Similarly, the transliteration words had a huge impact on common term count and, hence, on similarity computation. The number of common terms is directly proportional to the length of the Urdu news articles, and it was observed that five (05) transliterated words exist in the Urdu news articles. The results improved by 22%, because 75% of the common terms count increases, as is shown.

The SMTW similarity measure showed better performance than the CRMDL for linking dual-language news articles in the DNSA. It was observed that the SMTW performed well on large datasets (shown in Table 10). The study further concluded that sports news contained more English transliterated words in Urdu news articles and produced better results, and short Urdu news was hardly affected by transliteration words. The results improved by 20% (6 out of 30), dropped results by 04%, and 76% of the results remain unchanged. Urdu news articles contained about 20–30% transliterated words, depending on the type (Urdu and English) and length of news articles.

Table 9. Improved Results by SMTW Approach in 20 News Articles Set

	Ranked Results Muzi			**Transliteration Words**		
Urdu News	**CRMDL**	**SMTW**	**Results Impact**	**CRMDL**	**SMTW**	**CT Impact**
ur1	eng1	eng1	None	14	18	▲
	eng10	eng6	None	13	13	-
	eng6	eng10	None	11	13	▲
	eng7	eng7	-	18	20	▲
	eng4	eng4	-	4	6	▲
ur2	eng2	eng2	None	18	37	▲
	eng6	eng6	None	11	11	-
	eng8	eng8	None	12	12	▲
	eng5	eng5	-	9	9	-
	eng3	eng3	-	9	9	-
ur3	eng3	eng3	None	25	25	-
	eng4	eng4	-	12	12	-
ur4	eng4	eng4	None	26	41	▲
	eng7	eng7	-	15	15	-
ur5	eng5	eng5	None	21	37	▲
	eng1	eng1	-	11	16	▲
ur6	eng6	eng6	None	18	41	▲
	eng2	eng2	None	8	15	▲
	eng1	eng8	-	6	16	▲
	eng7	eng1	None	17	8	▲
	eng10	eng7	-	7	23	▲
	eng8	eng10	Dropped	4	9	▲
ur7	eng7	eng7	None	52	98	▲
	eng1	eng10	-	31	55	▲
	eng10	eng6	-	28	52	▲
	eng3	eng1	-	19	31	▲
	eng6	eng9	Improved	17	24	▲
ur8	eng8	eng8	None	22	38	▲
	eng3	eng6	Improved	8	19	▲
	eng4	eng2	Improved	8	12	▲
	eng1	eng4	-	4	13	▲
	eng2	eng3	-	4	8	▲
ur9	eng7	eng9	Improved	21	38	▲
	eng9	eng7	None	10	49	▲
	eng5	eng6	-	4	15	▲
ur10	eng1	eng10	Improved	17	42	▲
	eng10	eng6	None	20	24	▲
	eng6	eng1	None	13	17	▲
	eng7	eng7	-	21	33	▲

Figures 2 and 3 present the results of the precision and recall for all the datasets of news articles. The proposed similarity measure of the SMTW achieved more accurate and comprehensive results than the CRMDL for linking dual-language news articles in the DNSA.

Table 10. Results improvement by SMTW approach for one-day news article set, ▼ shows results impact is negative or dropped, ▲ shows results are improved and "-" represents "No Change or No impact".

	Ranked Results			Transliteration Words		
Eng News	**CRMDL Rank**	**SMTW Rank**	**Results Impact**	**CRMDL CT**	**SMTW CT**	**CT Impact**
UrNews	بجٹ 2017ـ18: سرکاری ملازمین کو خوش کردیا گیا					
Eng1	1	1	-	49	75	▲
Eng2	2	2	-	22	31	▲
Eng3	3	3	-	22	31	▲
Eng4	4	4	-	25	36	▲
Eng5	7	7	-	26	34	▲
Eng6	12	6	▲	12	20	▲
UrNews	رمضان المبارک کا چاند نظر آگیا، پہلا روزہ کل ہوگا					
Eng1	1	1	-	14	18	▲
Eng2	2	2	-	12	14	▲
Eng3	4	4	-	06	06	-
Eng4	9	9	-	04	04	-
Eng5	12	12	-	07	07	-
Eng6	13	13	-	06	06	-
Eng7	17	17	-	06	06	-
Eng8	18	19	-	11	11	-
Eng9	24	26	-	02	02	-
UrNews	بجٹ تنخواہ و پنشن میں 10 فیصد اضافے کی تجویز					
Eng1	1	1	-	82	121	▲
Eng2	2	2	-	83	115	▲
Eng3	3	3	-	162	219	▲
Eng4	4	6	-	87	106	▲
Eng5	5	4	-	66	97	▲
Eng6	6	8	-	55	83	▲
Eng7	7	7	-	56	86	▲
Eng8	8	5	-	42	71	▲
UrNews	یونس خان کی ''آل ٹائم'' ٹیسٹ ٹیم کے کپتان عمران خان					
Eng1	1	2	-	53	122	▲
Eng2	2	1	▼	65	176	▲
Eng3	6	4	▲	37	81	▲
Eng4	18	9	▲	27	69	▲
Eng5	26	3	▲	24	103	▲
Eng6	35	11	▲	19	50	▲
Eng7	51	7	▲	13	71	▲

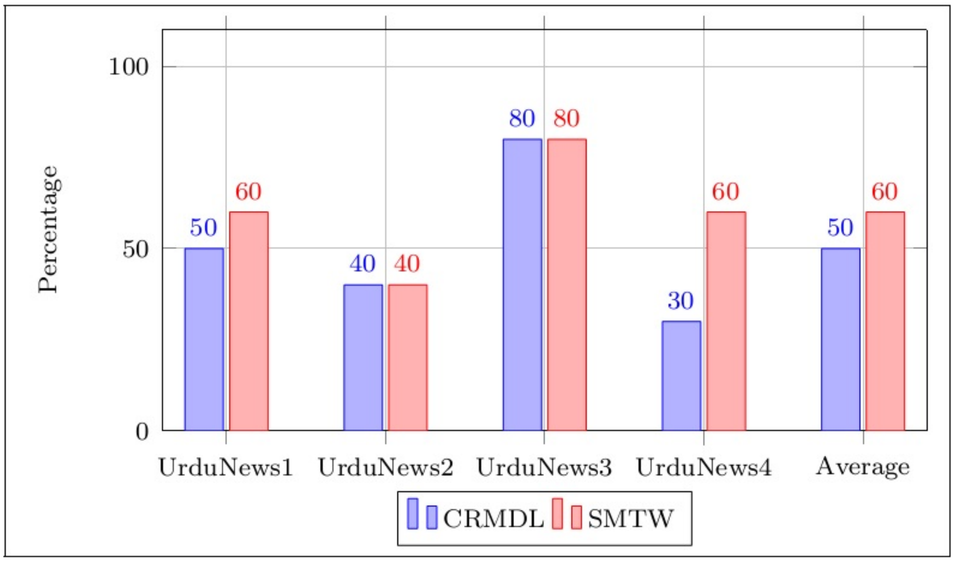

Figure 2. Precision comparison.

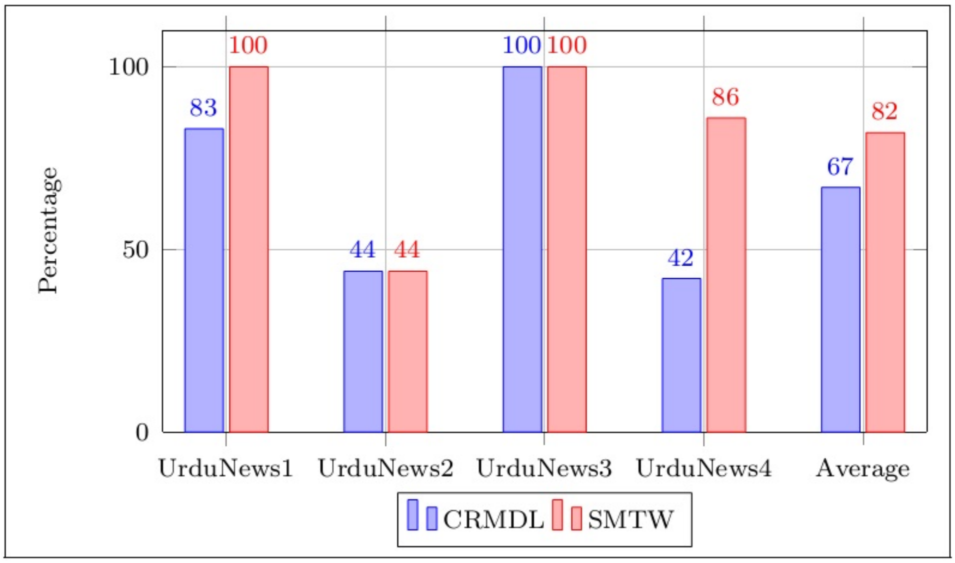

Figure 3. Recall comparison.

5. Conclusions and Future Work

The digital news preservation and management of low-resource languages are challenging tasks, especially for vast collections. The unique identification of individual digital objects is possible with well-defined attributes to assure efficient management, such as access, retrieval, preservation, usability, and transformability. The SMTW was introduced to utilize the transliteration words used in Urdu script for linking news articles during preservation to make it part of the metadata to manipulate and avoid run-time computation overhead. The proposed technique uses an Urdu-to-English lexicon for preprocessing enriched transliteration words. The analysis showed that about 9.5% of the transliteration words were contained in an Urdu script, thereby affecting the similarity value among news articles. The SMTW showed better results than the CRMDL technique, wherein it showed that 78% of Urdu news contained transliterated words. The precision improved from 50% to 60%, recall improved from 67% to 82%, and the impact of common terms also improved. The SMTW was effective and feasible for sports news. The extraction of Urdu news articles from diverse platforms and the consistent tokenization of Urdu manuscripts was one of the challenging tasks in the preprocessing step of the proposed lexical similarity approach. The results showed that the use of English transliteration words in Urdu scripts had a high impact in computing similarity to facilitate the linking of Urdu news articles with English news articles during preservation and archiving. The study made the following contributions:

- The DNSP framework was enhanced to a multilingual framework by including low-resourced languages, such as Urdu and Arabic.
- The study introduced a content-based approach for linking Urdu news articles to English news articles during preservation, i.e., it used a Similarity Measure based on Transliteration Words (SMTW).
- We designed a dataset to serve different purposes and steps of the evaluation.
- A comprehensive experiment was performed to assess the impact of English transliteration words that adopted both the user's centric and system-centric evaluation.
- The SMTW showed better results comparatively.
- The SMTW could generalize for other low-resource languages having the same character sets such as Arabic and Pashto languages.
- The main limitation of the Urdu and Arabic languages is the lack of availability of tools for tokenization and other preprocessing tasks. The Arabic and Pashto scripts need to be analyzed in more detail for the applicability of the SMTW.

The study presented details as to how the framework was enhanced and needs a more detailed study for accurate news content extraction and archiving for future access. The framework can be extended in different dimensions in the future, such as through the following improvements:

- The Arabic script needs to be analyzed in detail for multi-lingual linking.
- A standard user interface is required to enable access to the archived contents of the DNSA.
- The DNSE tool needs to be developed to a professional standard.
- The meta attributes can be developed for multi-lingual archives and other languages, such as Urdu, Arabic, Pashto, etc.
- Implicit meta elements can be added to the proposed set after comprehensively reviewing individual sources.
- We are working to improve the structure of the Urdu-to-English lexicon and the bag of Urdu words for efficient processing.
- More sophisticated content-based similarity measures need to be designed using different features, such as weighted terms, named entities, term position, and the context of the terms used in the news articles.
- The DNSA needs crossed-lingual techniques for linking multi-lingual archived news.

Author Contributions: M.K.: conceptualization, methodology, experimentation, development, data collection, and manuscript writing. S.S.K.: conceptualization, methodology, experimentation, manuscript writing, and proofreading. Y.A.: conceptualization, methodology, proofreading, and supervision. A.A.: conceptualization, proofreading, and supervision. T.S.A.: conceptualization, methodology, proofreading, and supervision. K.Y.: conceptualization, methodology, and proofreading. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Deanship at the University of Ha'il – Saudi Arabia, through project number RG-21090.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This article does not involve humans or animals.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SMTW	Similarity Measure based on Transliteration Words
CRMDL	Common Ratio Measure for Dual Language
WWW	World Wide Web
DNSA	Digital News Stories Archive
DNSP	Digital News Stories Preservation
DNSE	Digital News Stories Extractor
CT	Common Terms
TT	Total Terms
UT	Uncommon Terms
UrN	Urdu News
EngN	English News
Ur	Urdu
Eng	English
ICADL	International Conference on Asian Digital Libraries
AI	Artificial intelligence

References

1. Khan, M. Using Text Processing Techniques for Linking News Stories for Digital Preservation. Ph.D. Thesis, Faculty of Computer Science, Preston University Kohat, Islamabad Campus, HEC Pakistan, Kohat, Pakistan, 2018.
2. Grimes, C. Our New Search Index: Caffeine. 2010. Available online: https://developers.google.com/search/blog/2010/06/our-new-search-index-caffeine (accessed on 1 February 2023).
3. Size, W. The Size of the World Wide Web. 2022. Available online: https://www.worldwidewebsize.com/ (accessed on 4 August 2022).
4. Lavoie, B.F. The open archival information system reference model: Introductory guide. *Microform Digit. Rev.* **2004**, *33*, 68–81.
5. Guta, M. Small Business Trends. 15 February 2019. Available online: https://smallbiztrends.com/2019/02/web-hosting-uptime-statistics.html (accessed on 11 May 2022).
6. Burda, D.; Teuteberg, F. Sustaining accessibility of information through digital preservation: A literature review. *J. Inf. Sci.* **2013**, *39*, 442–458.
7. da Silva, J.R.; Ribeiro, C.; Lopes, J.C. A Data Curation Experiment at U. Porto using DSpace. In Proceedings of the 8th International Conference on Preservation of Digital Objects, Singapore, 1–4 November 2011.
8. Rahman, A.U.; David, G.; Ribeiro, C. Model migration approach for database preservation. In Proceedings of the International Conference on Asian Digital Libraries, Gold Coast, Australia, 21–25 June 2010; pp. 81–90.
9. Khan, M.; Alharbi, Y.; Alferaidi, A.; Saad, A.T.; Yadav, K. Metadata for Efficient Management of Digital News Articles in Multilingual News Archives. *SAGE Open* **2023**, *13*, 1–17.
10. Dashrath, V.B. Role of metadata in digital resource management. *Int. J. Digit. Libr. Serv.* **2014**, *4*, 209–2017.
11. Hajiyev, A. Artificial Intelligence in the Newsroom. In *Mass Communication*; Liberty Academic Publishers: New York, NY, USA, 2022; pp. 68–71.
12. Khan, M.; Rahman, A.U. Digital News Story Preservation Framework. In Proceedings of the Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, 9–12 December 2015; Volume 9469, p. 350.
13. Khan, M.; Rahman, A.U.; Awan, M.D.; Alam, S.M. Normalizing digital news-stories for preservation. In Proceedings of the Digital Information Management (ICDIM), 2016 Eleventh International Conference, Porto, Portugal, 19–21 September 2016; pp. 85–90.
14. Khan, M.; Rahman, A.U. A Systematic Approach Towards Web Preservation. *Inf. Technol. Libr.* **2019**, *38*, 71–90.
15. Khan, M.; Rahman, A.U.; Awan, M.D. Exploring the Digital World of Newspaper Archives. *Sci. Technol. J. Port.* **2017**, *32*, 140–164.
16. Khan, M.; Rahman, A.U.; Awan, M.D. Term-Based Approach for Linking Digital News Stories. In Proceedings of the Italian Research Conference on Digital Libraries, Udine, Italy, 25–26 January 2018; pp. 127–138.
17. Khan, M.; Rahman, A.U.; Ullah, M.; Naseem, R. The Role of Named Entities in Linking News Articles During Preservation. In Proceedings of the International Conference on the Sciences of Electronics, Technologies of Information and Telecommunications, Genoa, Italy, 18–20 December 2018; pp. 50–58.
18. Feng, C.; Khan, M.; Rahman, A.U.; Ahmad, A. News Recommendation Systems-Accomplishments, Challenges & Future Directions. *IEEE Access* **2020**, *8*, 16702–16725.
19. Khan, M.; Rahman, A.U.; Ahmad, A.; Khan, S.S. A content-based technique for linking dual language news articles in an archive. *J. Inf. Sci.* **2020**, *48*, 57–70.
20. Khan, M.; Khan, S.S.; Ahmad, A.; Rahman, A.U. The role of news title for linking during preservation process in digital archives. *Libr. Hi Tech* **2020**, *40*, 1359–1383.

21. Athalye, S. Recommendation System for News Reader. Master's Thesis, San Jose State University, San Jose, CA, USA, 2013.
22. Melville, P.; Sindhwani, V. Recommender systems. *Encycl. Mach. Learn.* **2011**, *1*, 829–838.
23. Doychev, D.; Lawlor, A.; Rafter, R.; Smyth, B. An Analysis of Recommender Algorithms for Online News. In Proceedings of the CLEF (Working Notes), Sheffield, UK, 15–18 September 2014; pp. 825–836.
24. Kutsuki, A. Do bilinguals acquire similar words to monolinguals? An examination of word acquisition and the similarity effect in japanese—English bilinguals' vocabularies. *Eur. J. Investig. Health Psychol. Educ.* **2021**, *11*, 168–182.
25. Agarwal, D.; Chen, B.C.; Elango, P.; Wang, X. Personalized click shaping through lagrangian duality for online recommendation. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, USA, 12–16 August 2012; pp. 485–494.
26. Fortuna, B.; Fortuna, C.; Mladenić, D. Real-time news recommender system. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, 20–24 September 2010; pp. 583–586.
27. Li, L.; Wang, D.D.; Zhu, S.Z.; Li, T. Personalized news recommendation: A review and an experimental investigation. *J. Comput. Sci. Technol.* **2011**, *26*, 754–766.
28. Li, L.; Zheng, L.; Yang, F.; Li, T. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Syst. Appl.* **2014**, *41*, 3168–3177.
29. Said, A.; Bellogín, A.; Lin, J.; de Vries, A. Do recommendations matter? News recommendation in real life. In Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Portland, OR, USA, 25 February–1 March 2017; pp. 237–240.
30. Li, L.; Li, T. News recommendation via hypergraph learning: Encapsulation of user behavior and news content. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 305–314.
31. Borrow Language Definition. Available online: https://www.thoughtco.com/what-is-borrowing-language-1689176 (accessed on 5 January 2023).
32. Accredited Language Services. Available online: https://www.accreditedlanguage.com/2016/09/09/what-is-transliteration/ (accessed on 5 January 2023).
33. Al-Onaizan, Y.; Knight, K. Machine transliteration of names in Arabic text. In Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA, 11 July 2002; pp. 1–13.
34. Alam, S.M.; Rehman, A.U.; Khan, M. Quantifying the Use of English Words in Urdu News-Stories. In Proceedings of the Student Conference on Engineering Sciences and Technology, SCONEST, Karachi, Pakistan, 14–15 December 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.