Engineering Applications of Artificial Intelligence 125 (2023) 106718

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Improving news headline text generation quality through frequent POS-Tag patterns analysis

Noureen Fatima [a], Sher Muhammad Daudpota [a], Zenun Kastrati [b], Ali Shariq Imran [c,*], Saif Hassan [a], Nouh Sabri Elmitwally [d]

[a] *Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan*
[b] *Department of Informatics, Linnaeus University, 35195 Växjö, Sweden*
[c] *Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway*
[d] *School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK*

ARTICLE INFO

Keywords:
POS tagging
Text generation
Low resource language
Generative pre-trained transformer
Attention mechanism

ABSTRACT

Original synthetic content writing is one of the human abilities that algorithms aspire to emulate. The advent of sophisticated algorithms, especially based on neural networks has shown promising results in recent times. A watershed moment was witnessed when the attention mechanism was introduced which paved the way for transformers, a new exciting architecture in natural language processing. Recent sensations like GPT and BERT for synthetic text generation rely on NLP transformers. Although, GPT and BERT-based models are capable of generating creative text given they are properly trained on abundant data, however, the generated text suffers the quality aspect when limited data is available. This is especially an issue for low-resource languages where labeled data is still scarce. In such cases, the generated text, more often than not, lacks the proper sentence structure, thus unreadable. This study proposes a post-processing step in text generation that improves the quality of generated text through the GPT model. The proposed post-processing step is based on the analysis of POS tagging patterns in the original text and accepts only those generated sentences from GPT which satisfy POS patterns that are originally learned from the data. We exploit the GPT model to generate English headlines by utilizing Australian Broadcasting Corporation (ABC) news dataset. Furthermore, for assessing the applicability of the model in low-resource languages, we also train the model on the Urdu news dataset for Urdu news headlines generation. The experiments presented in this paper on these datasets from high- and low-resource languages show that the performance of generated headlines has a significant improvement by using the proposed headline POS pattern extraction. We evaluate the performance through subjective evaluation as well as using text generation quality metrics like BLEU and ROUGE.

1. Introduction

It is a long-held belief among Artificial Intelligence (AI) community that AI is not only for passive tasks such as object detection, classification, or clustering of input instances based on similarity, rather AI is destined to emulate much high-order human endowment such as creative content writing, original music composing, Picasso-level painting generation, etc. The great Alan Turing once said that a computer can be said to possess artificial intelligence if it can mimic human responses under specific conditions (Turing, 2009).

Content writing apparently seems to be an easy writing task in which some repetitive linguistic steps are followed to convey a message. However, it is quite the opposite of that. It is a creative process whereby good writers avoid cliches and monotonous expressions. Rather, they come up with innovative and interesting ways of expressing the content that attracts the readers to read and remember the core message. In its best form, it stimulates the readers to take action. No doubt, it does not mean that there are not any predictable patterns and principles in writing the best content. Content writing is a precious human ability that only recently AI through deep learning models has started to mimic. Significant changes have been made with the development of deep learning models in the field of natural language processing (Imran et al., 2020), speech (Fatima et al., 2022b), and image (Lv et al., 2021). In addition to this, deep learning-based models could be used to extract those principles and generate original and captivating messages. In recent years, different aspects of content writing have become hot topics in the field of natural language processing including news generation (Nishi et al., 2021), poetry writing (Talafha and Rekabdar, 2021), headline generation (Shen et al., 2017), etc. This is possible

* Corresponding author.
E-mail addresses: noureen.mscss19@iba-suk.edu.pk (N. Fatima), sher@iba-suk.edu.pk (S.M. Daudpota), zenun.kastrati@lnu.se (Z. Kastrati), ali.imran@ntnu.no (A.S. Imran), saif.hassan@iba-suk.edu.pk (S. Hassan), nouh.elmitwally@bcu.ac.uk (N.S. Elmitwally).

https://doi.org/10.1016/j.engappai.2023.106718
Received 29 August 2022; Received in revised form 11 June 2023; Accepted 23 June 2023
Available online xxxx

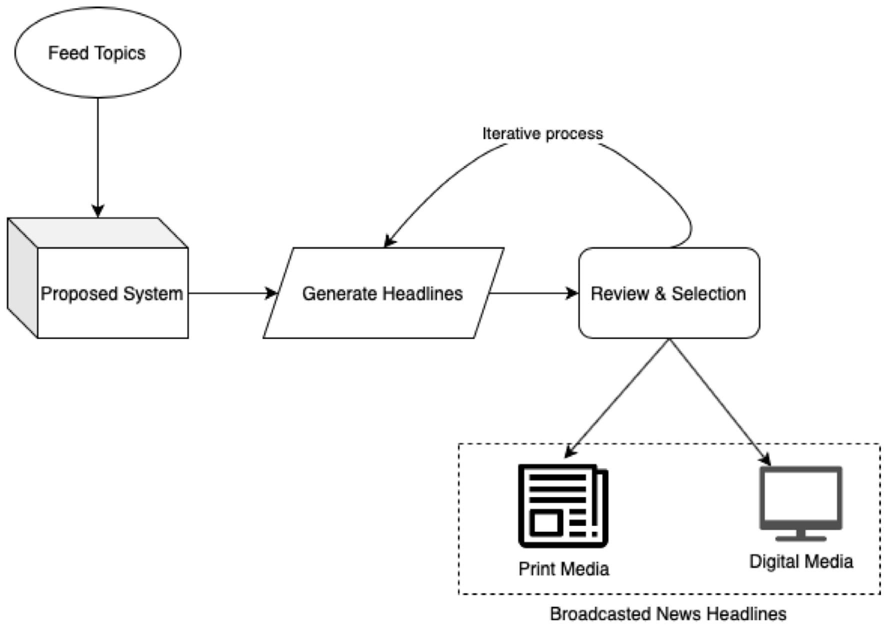

Fig. 1. Application usage of the proposed system.

after having advanced deep learning generation models like GPT-2 (Li et al., 2021; Shaikh et al., 2021) and GPT-3 (Shavrina and Shliazhko, 2021) and GAN based models (Haidar et al., 2019; Liu et al., 2020; Imran et al., 2022). However, the GPT-2 model has several advantages compared to GAN. First, GPT-2 uses a powerful deep learning network with 1.5 billion parameters, making it capable of producing much higher quality results than GAN (Cao et al., 2023). Because of this, GPT-2 is better suited for short-text generation tasks such as news headlines or summarization (Zhu and Luo, 2023). Additionally, GPT-2 can be trained faster than GPT-3 since its architecture is simpler, making it the most cost-effective model for the task at hand (Kolides et al., 2023; Xu et al., 2022).

News headlines generation is one of the text generation tasks which expects to generate a small text snippet that summarizes larger news contents. It assists in indexing the news as well as attracting readers to read full news stories. News headlines are not only short summaries of the news but need to be an eye-catching statements for the newsreader. Many researchers have proposed headline generation models that generate human-like headline scripts (Shenassa and Minaei-Bidgoli, 2022). These generated headlines can be evaluated through human/language experts or objectively using machine-centric approaches like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and a different variant of these metrics like BLEU (1-unigram, 2-bigrams, 3-trigram to n-gram) and ROUGH (1-unigram, 2-bigrams, L-Longest Common Subsequence, W-Weighted, S-Skip-bigram based co-occurrence statistics and SU-Skip-bigram plus unigram-based co-occurrence).

In addition to this, headline news generation has been explored in many languages for instance, English (Barros et al., 2021), Japanese (Tseng et al., 2022), Chinese (Wang et al., 2020), Persian (Shenassa and Minaei-Bidgoli, 2022), and Russian (Gavrilov et al., 2019). To the best of our knowledge, no one has yet worked on synthetic news headline generation for low-resource languages like Urdu.

Although recent developments in deep learning have enabled AI to generate synthetic text, however, these require abundant labeled data which is scarce in low-resource languages. Resultantly, the models like GPTs and GANs mostly generate low-quality text if trained on a smaller dataset. The generated text is of low quality because it does not follow the sentence structure in that language. Therefore in this study, we add a post-processing step in the GPT model to further improve generated text quality by ensuring a correct sentence structure of the generated text.

1.1. Research questions and objectives

This research study proposes a synthetic news analysis and synthesis for low-resource languages along with applying the same model for high-resource language i.e. English to aid print and digital news media companies, as shown in Fig. 1. The media outlets can generate captivating news automatically to be used via our proposed model presented in Section 3. Given the initial seed words (topics), the presented trained model can generate synthetically and semantically appealing headlines that the media outlets can review and select to use. More specifically, we intend to address the following research questions (RQs):

1. RQ1: What is the impact of extracting the POS tag patterns from the original headline on generating the synthetic headline news?
2. RQ2: How accurate is the GPT-2 model in generating the headline news for both low-resource language (Urdu) and high-resource language (English)?
3. RQ3: What is the difference in performance between generated headlines evaluated through Turing tests and machine-centric approaches?

To address the research questions (RQ1–RQ3), we have performed several experiments on one Million Australian Broadcasting Corporation (ABC) news dataset for the English language and one Million Urdu News Dataset for the Urdu language. We have employed GPT-2 model for headline generation. To improve the quality of the generated headlines, we have performed post-processing techniques which are based on extracting POS patterns from the original news dataset. A sentence is composed of different POS tags such as a noun, preposition, adjective, verb, conjunction, etc. The sequence in which they appear in the language governs a language structure. Fig. 2 shows a sample of dataset sentences with POS patterns. We exploit these patterns in our dataset and extract the frequently occurring patterns to learn the correct structure of a sentence for generating synthetic headlines. The quality assessment of the generated news headlines is carried out through subjective evaluation using Turing Test as well as objective evaluation using different text generation quality metrics.

The Turing Test was initiated by Alan Turing (Hodges, 2014). The main concept behind this test is to determine whether or not

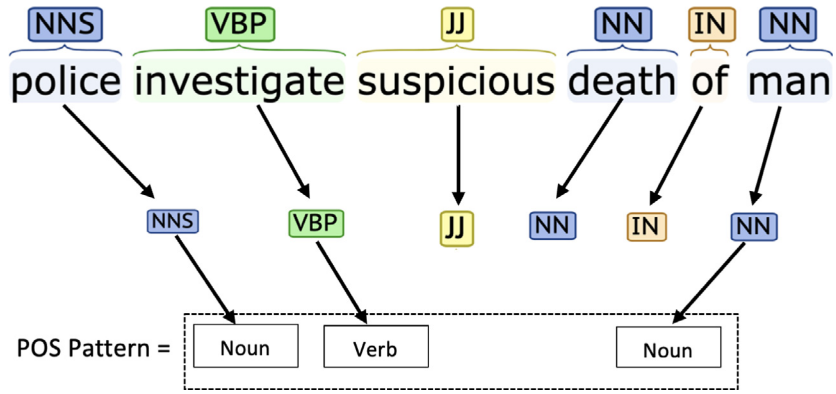

Fig. 2. An example of a POS pattern.

a machine is capable of generating original creative content like a human being. Thus, to evaluate generated text, we conduct a human evaluation and ask explicitly from a different group of people whether headlines are generated by humans or machines. An extensive set of experimental analyses was performed to evaluate the performance of our proposed approach. The experimental results show that our POS tagging-based approach has brought significant improvement in the generated text when evaluated subjectively through Turing Test and objectively through machine-based quality metrics.

1.2. Problem statement

Let H denote a set of original headlines and P denote a set of POS patterns extracted from H. The objective of this research is to propose a synthetic headline news generation model for both low-resource language (Urdu) and high-resource language (English) by using GPT-2. Moreover, let G denote the set of generated headlines by the proposed model when the model is trained on H. The performance of the generated headlines is evaluated through both the Turing test (human subjective evaluation) and machine-centric approaches.

The problem can be formulated mathematically as follows:

Given a set of original headlines,

$$H = h_1, h_2, \dots, h_n \tag{1}$$

the set of extracted POS patterns which is extracted from H is defined as,

$$P = p_1, p_2, \dots, p_m \tag{2}$$

a synthetic headline news generation model GPT-2, which is trained on H, the objective is to generate a set of headlines,

$$G = g_1, g_2, \dots, g_k \tag{3}$$

for both low-resource language (Urdu) and high-resource language (English) that are syntactically and semantically coherent and align with the POS patterns P extracted from H.

1.3. Contributions

The main contributions of this study are as follows:

1. Extracted POS pattern from original headlines to validate the generated headlines.
2. Proposed a synthetic headline news generation model by using GPT-2 for both low-resource language (Urdu) and high-resource language (English).
3. Evaluated the performance of generated headlines through both the Turing test (human subjective evaluation) and machine-centric approaches.

The rest of the article is structured as follows. Section 2 presents the related work followed by a methodology in Section 3. Results and detailed discussion are presented in Section 4. Lastly, Section 5 concludes the paper along with some future directions for new researchers.

2. Related work

Headline news generation has been a topic of interest in recent years for many researchers. Recent studies have shown promising results on headline news generation task by employing deep learning models like RNN (Rumelhart et al., 1994), GRU (Chung et al., 2014), CNN (O'Shea and Nash, 2015) and LSTM (Schmidhuber, 2015). To represent the semantic better, attention-based transformer architecture (Vaswani et al., 2017) outperforms previous recurrent models. Table 1 shows important research studies which are conducted on headlines generation tasks. Several researchers have employed reinforcement models to generate syntactic headline news.

The study conducted in Barros et al. (2021) proposed a hybrid surface natural language generation (HanaNLG) technique to generate the headline through the text summarization technique. The authors performed text generation in three steps; prepossessing, macro & micro-planning, and surface realization. In the prepossessing step, they analyzed (lexical, syntactic, and semantic) using a language analyzer. In the second step, they created the vocabulary that will be used for generating the final headlines through the different methods; Name Entity (NE), Latent Dirichlet Allocation (LDA), Term Frequency-Inverse Sentence Frequency (TF-ISF), and Positional Language Models (PLM). They have evaluated the model on each generating the vocab technique. In last, they applied the surface realization step where they generated the headline through Sentence Generation and Sentence Ranking algorithms. They have tested their proposed method on DUC 2003 and DUC 2004 datasets. To evaluate the generated text, they have used both human and machine-centric approaches. After analyzing the results, they stated that the TF and PLM strategy outperforms both the human and machine-centric approaches.

Many researchers have used RNN, CNN, LSTM, and GRU to generate headline news generation. For example, in the research work conducted in Fujita and Watanabe (2018), CNN and LSTM were employed to generate Japanese news headlines. The authors have utilized CNN/Daily Mail, Gigaword, and NEWSROOM benchmark datasets to generate the syntactic news. Generated news were evaluated through the questionnaires method. Similarly, the researchers used the topic modeling along with RNN and GRU to generate Chinese news headlines in the 2020 study (Wang et al., 2020). They have proposed a topic-sensitive neural headline generation model that can generate headline news by utilizing the Chinese short text summarization dataset.

In recent work, two of the studies (Xie et al., 2019; Shen et al., 2018) have used the encoder–decoder architecture to generate the headline news. For instance, researchers in Xie et al. (2019) have generated

Table 1
Related work on headlines generation.

Ref	Year	Approach	Dataset	Evaluation	Language
Tseng et al. (2022)	2022	Transformer-based models	Mai-news	Machine centric	Japanese
Shenassa and Minaei-Bidgoli (2022)	2022	Transformer-based and LSTM	LSTM and transformer-based Hamshahri (contains 166000 news headline pairs) and Persica and Tabnak Agency (contains 10000 news–headline)	Machine	Persian
Barros et al. (2021)	2021	NE, TF-ISF, LDA, TF, and PLM	DUC 2003 and DUC 2004	Both	English
Singh et al. (2021)	2021	Markov model	Daily Mail dataset, Gigaword dataset, NEWSROOM dataset	Machine centric	English
Mishra and Zhang (2021)	2021	bi-Lstm with attention	NELA17	N/A	English
Li et al. (2021)	2021	RNN, LSTM, and Transformer	large corpus of Chinese short text summarization	Machine centric	Chinese
Wang et al. (2020)	2020	RNN and GRU(topic modeling)	large corpus of Chinese short text summarization	Both	Chinese
Chen et al. (2020b)	2020	Reinforcement learning	LDC2002E18, LDC2003E07, LDC2003E14, and part of LDC2004T07, LDC2004T08, and LDC2005T06	Machine-centric	English and Chines
Xie et al. (2019)	2019	Encoder–Decoder Attention	English gigaword, DUC, Abstractive Text Compression(MSR)	Machine centric	English
Shen et al. (2018)	2018	Encoder–Decoder	DUC 2003, DUC 2004, LDC2002E18 LDC2003E07, LDC2003E14, LDC2004T08, and LDC2005T06	Machine-centric	English and Chines
Fujita and Watanabe (2018)	2018	CNN and LSTM	CNN/Daily Mail, Gigaword, and NEWS-ROOM	Human Centric	Japanese
Alfonseca et al. (2013)	2013	Clustering Approach	Private dataset	Both	English

the English abstract headlines by testing three well-known datasets namely, the English gigaword, DUC, and abstractive text compression (MSR) dataset. The main objective of this paper was to generate an abstractive headline based on an eye-tracking attention mechanism. In eye-tracking, the main idea was to construct a method in which the importance of different words is being extracted. Overall, they obtained significant results by evaluating the generated text on ROUGE-1, ROUGE-2, and ROUGE-L quality metrics. Apart from this, similar architecture was employed to generate the headline for the multilingual task in research work conducted in Shen et al. (2018). The authors have generated headline news in Chinese and English.

Headline news generation using transformers has shown a lot of interest among researchers in recent years. For instance, the study (Tseng et al., 2022) proposes to generate Japanese news headlines through a transformer-based approach by using the Mia news dataset. The finding showed that transformer-based approach was able to generate the synthetic Japanese news headline. Similarly, Shenassa and Minaei-Bidgoli (2022) have worked on Persian news headline generation through BiLSTM and transformer generation models. Two benchmark datasets were utilized to generate the headlines, namely Hamshahri (contains 166,000 news–headline pairs) and Persian and Tabnak Agency (contains 10,000 news–headline). To evaluate the generated headlines, the ROUGE metric was used. A multilingual translation (Chinese to English) approach for generating headlines was proposed by Singh and Josan (2022). In this approach, the authors have trained the models on various datasets namely, DUC 2003, DUC 2004, LDC2002E18, LDC2003E07, LDC2003E14, LDC2004T07, LDC2004T08, and LDC2005T06, and validated the model by generating their own dataset for English-Chinese cross-lingual headline generation. To evaluate the generated headlines, they used BLEU and ROUGE-1, ROUGE-2, and ROUGE-L. Later, this multilingual approach was extended by the same authors by employing reinforcement learning techniques to generate the news headline (Chen et al., 2020b). Likewise, in another study in Singh et al. (2021), the authors have performed headline generation through reinforcement learning models.

Similarly, very few researchers have worked on extracting the syntactic and ontological information through POS patterns to generate the synthetic news headline on high-resource language(English) (Alfonseca et al., 2013; Mosallanezhad et al., 2020, 2021). For instance, a clustering-based approach was proposed by this study (Alfonseca et al., 2013) in which, they worked on English news headlines by extracting the syntactic and ontological information by using the Bayesian network. In this study (Mishra and Zhang, 2021), a topic-preserving synthetic news generation was introduced in which a reinforcement learning agent was employed to select words (rather than any text generation model) that optimizes the matching of a given topic. Also, these studies (Lin et al., 2021, 2020, 2019; Shao et al., 2021; Pal et al., 2022; Mishra and Zhang, 2021) have proven that POS taggers perform better and are being implemented as potential solutions to efficiently identify patterns in any given domain like summary generation (Pal et al., 2022), labeling task (Lin et al., 2021, 2020, 2019; Shao et al., 2021; Pal et al., 2022; Mishra and Zhang, 2021).

Recently, a systematic literature review paper on text generation is conducted by Fatima et al. (2022a). The authors have conducted an extensive review of text generation in five aspects mainly focused on deep learning approaches, quality metrics, datasets, languages, and applications from 2015 to 2021. One of the aims of this article was to find the most studied languages for which text-generation techniques are utilized. They have found that only 23% of researchers have worked on low- to mid-resource languages like Arabic (Hejazi et al., 2021), Spanish (Wang and Issa, 2020), Turkish (Chen et al., 2020a), and many others. The authors emphasized the need to explore low-resource languages with respect to building and training models for synthetic text

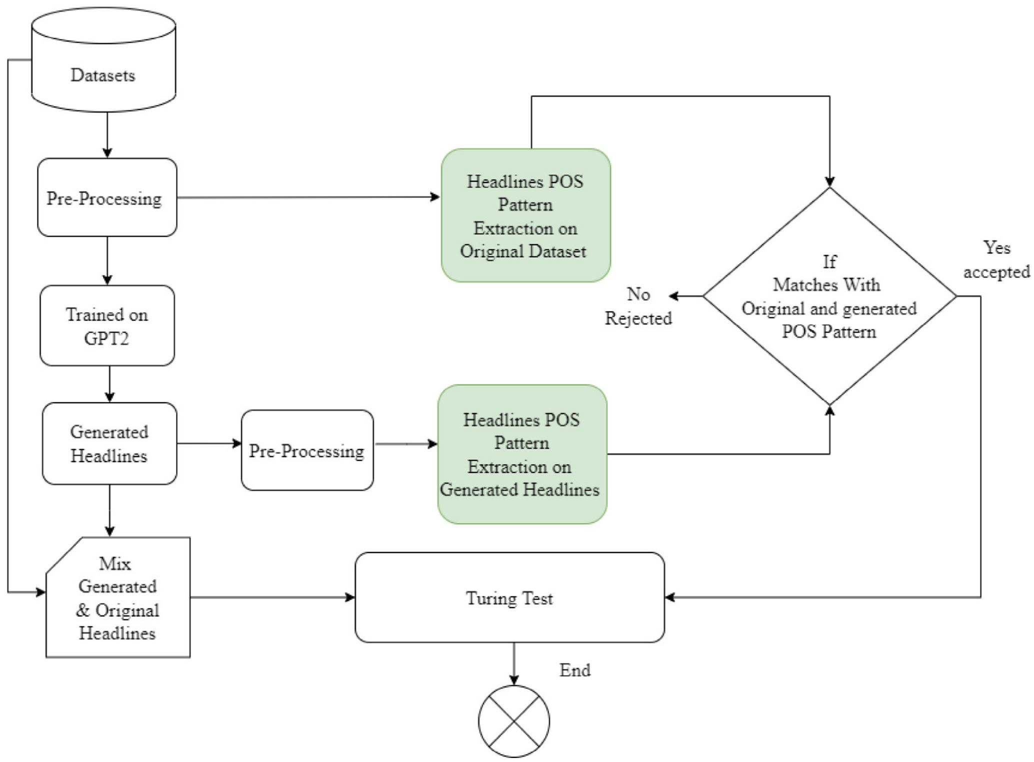

Fig. 3. Abstract model of the proposed system.

generation. Although the low-resource languages have been utilized in many other tasks like sentiment analysis (Chandio et al., 2022), classification and categorization (Hossain et al., 2021; Batra et al., 2021; Hossain et al., 2023), emotion detection (Ashraf et al., 2022), translation (Ghafoor et al., 2021), there is a lack of research works in text generation.

Our study is different from the aforementioned approaches in three aspects. First, to the best of our knowledge, it is the first study that attempts to work on synthetic news headline analysis for the low-resource Urdu language and high-resource English language combined. More concisely, we have used the one Million Australian Broadcasting Corporation (ABC) news dataset for the English language and one Million Urdu news dataset for the Urdu language. We have employed GPT-2 model for headline generation. Second, we introduce a POS-tagging-based post-processing step in text generation which improves the quality of the generated text. The approach is generalizable regardless of the language or model being exploited. Third, an in-depth performance analysis using a Turing test involving domain expert skills to assess the generated headline quality along with quality metrics for text generation is performed.

3. Methodology

The abstract model of the proposed system is presented in Fig. 3. It is composed of seven steps that are summarized below:

1. Perform pre-processing on the original dataset. In this step, we remove special characters, numbers, and punctuation. We do not remove stop words as these are required to be included in the generated headline.
2. Learn POS tags of all the instances in the input dataset. More than 3000 different POS tag combinations have been observed in both English and Urdu datasets, however, the majority of the POS tags are associated with only a few input instances. Whereas there are POS tag patterns that are more frequent, we retain those frequent POS tags and use them to validate the generated headline.
3. Train a GPT-2 model on the input dataset.
4. Generate the news headlines on the trained GPT-2 model.
5. Repeat steps one and two on the generated headlines.
6. Compare the POS pattern of the generated headline with POS pattern of the original headline. If the generated headline satisfies at least one POS tag pattern among the frequent POS tag patterns of the original headline, retain it for subjective and objective evaluation, else discard it.
7. Mix the generated headlines with original headlines and perform the Turing test for subjective evaluation and also use BLEU and ROUGE for objective evaluation.

In subsequent sections, we will explain each of these steps in more detail.

3.1. Dataset for low and high resource languages

In this work, we have experimented on two datasets, the details about each dataset are given below.

Low-Resource Language: The Urdu News dataset used in this study contains news from the major Urdu news sources such as 92 News,[1] Ab Tak News,[2] Dawn News,[3] Express News[4] and Geo News[5] which are published news in different categories (Hussain et al., 2021). A customized separate Python script using BeautifulSoup and Request libraries was used for data extraction from each category for each website. The pre-processing techniques employed use customized functions and regular expressions in Python to keep Urdu text and numbers only in the dataset corpus. This dataset has more than one million Urdu news stories text corpus for four distinct categories: Business & Economics, Science & Technology, Entertainment, and Sports from the year 2012 to 2020. The dataset is available online as open-source.[6]

1 https://urdu.92newshd.tv/
2 https://urdu.abbtakk.tv/
3 https://www.dawnnews.tv/
4 https://www.express.pk/
5 https://urdu.geo.tv/
6 https://data.mendeley.com/datasets/834vsxnb99/3

High-Resource Language: The English news dataset used in this study contains news from the Australian Broadcasting Corporation (ABC)[7] (Kulkarni, 2018). This dataset contains more than one million English news stories as a summarized historical record of noteworthy events at global levels with a more granular focus on Australia such as the Afghanistan war, financial crisis, multiple elections, ecological disasters, terrorism, famous people, and criminal activity from the year 2003 to 2020. The dataset is available online as open-source "A Million News Headlines".[8]

3.2. Pre-processing technique

While developing Urdu News Dataset and cleaning the ABC dataset, the following steps were followed in pre-processing:

- **Removed duplicates:** In the process of scraping news from online sources, duplicate news was also included. We excluded such items as these do not contribute to the richness of the overall dataset text.
- **Removed Non-Urdu Text:** There were news stories that mixed non-Urdu text with the news story which certainly does not contribute to text generation, therefore such text portions were removed which were in other languages.
- **Removed Hashtag, URLs, and white space:** has these embedded inside news text and do not contribute much to the text generation, therefore removed from the text.
- **Removed null values:** In some cases, after performing the above steps, there was nothing left in the instance, therefore such instances were also removed. For example, some of the news pages have only video links so the removal of that link results in null values in the instance which is eliminated in this step.

3.3. GPT-2 model for headline generation

Generative Pre-trained Transformer (GPT-2), proposed by Radford et al. (2019) and designed by OpenAI, is a transformer-based model having 1.5 billion parameters. It is trained on 40 GB of Internet text scrapped from eight million web pages. It is a revolutionary model in text processing. It has an exceptional human-like ability to generate long sequences. An article generated through GPT-3[9] was also published in The Guardian and became an internet sensation as well as a heated topic of discussion on electronic media. In addition to this, the latest version of GPT-2 is open source and capable of generating text for low-resource languages like Urdu, Arabic, and many others. The detailed summary of GPT-2 used for headline generation is given in Fig. 4.

3.3.1. GPT-2 customization for news headlines generation

In order to generate news headlines, we have done different corpus settings for GPT-2 as discussed below.

Corpus Setting for Low-Resource Language Dataset: The 1 million Urdu news dataset comprises three columns including the headline, complete news, and its label for class. We have utilized both columns' news and its headline in this experiment. We have done two settings, first for the headline and another for the news. For headlines, we have created the corpus for an entire dataset that contains all headlines related to all four classes. Similarly, we have also created the corpus for an entire dataset that contains news related to all four classes.

Corpus Setting for High-Resource Language Dataset: Similarly, for the English news headline, we have chosen the entire dataset as a single corpus that contains headlines related to all four classes.

[7] https://www.abc.net.au/

[8] https://www.kaggle.com/therohk/million-headlines

[9] https://www.theguardian.com/commentisfree/2020/sep/12/human-wrote-this-article-gpt-3

```
==== Embedding Layer ====

transformer.wte.weight                    (50259, 768)
transformer.wpe.weight                    (1024, 768)

==== First Transformer ====

transformer.h.0.ln_1.weight               (768,)
transformer.h.0.ln_1.bias                 (768,)
transformer.h.0.attn.c_attn.weight        (768, 2304)
transformer.h.0.attn.c_attn.bias          (2304,)
transformer.h.0.attn.c_proj.weight        (768, 768)
transformer.h.0.attn.c_proj.bias          (768,)
transformer.h.0.ln_2.weight               (768,)
transformer.h.0.ln_2.bias                 (768,)
transformer.h.0.mlp.c_fc.weight           (768, 3072)
transformer.h.0.mlp.c_fc.bias             (3072,)
transformer.h.0.mlp.c_proj.weight         (3072, 768)
transformer.h.0.mlp.c_proj.bias           (768,)

==== Output Layer ====

transformer.ln_f.weight                   (768,)
transformer.ln_f.bias                     (768,)
```

Fig. 4. GPT-2 model architecture.

3.3.2. Training of GPT-2 for news headlines generations

We start the training of the GPT-2 model using different corpus individually as explained in the previous step. Four main parameters are required in GPT-2 to start training the model: batch size, learning rate, warmup_steps, and sample_every. The values selected for these parameters for GPT-2 model for headline generation are shown below:

- learning_rate = 5e-4
- sample_every = 100
- warmup_steps = 1e-2
- batch size = 4

These values have been selected experimentally by observing different combinations. In addition to this, we have used stop criteria for training as learning loss_value = 0.01. The model for low- and high-resource language is the same, as shown in Fig. 4.

3.3.3. Generating new samples of headlines

Table 2 and Fig. 5 show a sample of the original news headline, before applying Headlines POS Pattern Extraction (HPPE) and after applying HPPE generated sample, where the detail of HPPE is discussed in Section 3.4. It can be observed that for both low and high-resource languages, the generated headlines are meaningful and very close to the real headline news.

3.4. Headlines POS pattern extraction (HPPE)

The experiments were conducted to produce text of different lengths, including 80, 100, and 120 characters. We tested various sizes for generating headlines and found that 120 characters were the optimal length. Finally, we compared the generated headlines of different sizes with the top 25 most common headlines, using their respective parts of speech (POS) tags, and found that the headlines generated with 120 characters were the most similar to the top 25 patterns. Consequently, it was decided to limit the generated text to 120 characters. Experimental results samples are shown in Figs. 6, 7 and 8.

In order to make these generated headlines more readable and synthetic, we performed post-processing. In the post-processing steps, first, we extracted all the POS tags from each headline for the original dataset by using Stanza library (Qi et al., 2020) for Urdu language POS Tagger as shown in Fig. 9. For the English language, we used the NLTK library. After extracting POS tags from actual headlines for both English and Urdu, we have extracted POS tags from GPT2-Generated text. The

Original Headlines	Sample Generated Headlines Before HPPE	Sample Generated Headlines After HPPE	
نیلم جہلم ہائیڈرو پاور پراجیکٹ شدید مالی بحران کا شکار Neelum Jhelum Hydropower Project is suffering from severe financial crisis	دبئی سپورٹس ڈیسک انٹرنیشنل کرکٹ کون Dubai Sports Desk International Cricket Con	گوگل نے مختلف حلقوں میں نئے فیچرز متعارف کروا دیے ہیں Google has introduced new features in various circles	1
فیس بک میسنجر پر کووڈ19 چیٹ بوٹ لانچ کر دیا گیا A Covid-19 chatbot has been launched on Facebook Messenger	فیس بک کی جانب سے اکثر نئی سیکیورٹی فیچرز شامل کیے جاتے ہیں Facebook often adds new security features	ڈسکوز کا بجلی کے ٹیرف میں 86 پیسے اضافے کا مطالبہ Discos demand 86 paisa increase in electricity tariff	2
اینٹی کورونا وائرس ٹیکنالوجی پر مبنی گاڑی متعارف Introducing a vehicle based on anti-coronavirus technology	سوشل میڈیا پر اپنے خیالات کے اظہار کا سلسلہ جاری ہے The series of expressing their views on social media continues	جوہری بجلی گھر کی زنائش کیلئے ایندھن کی لوڈنگ کا غاز Fuel loading gas for nuclear power plant loading	3
انڈیا سے لندن پہنچنے والا غریب بیلے ڈانسر A poor ballet dancer arriving in London from India	لاہور ئی این پیسابق اولمپینز نے چیف سلیکٹر کے عہدے سے استعفی دے Lahore ENP former Olympians resigned from the post of Chief Selector	پیٹرول کی قیمت برقرار ڈیزل روپے فی لیٹر مہنگا Petrol price unchanged, diesel expensive Rs. 1 per liter	4
پاکستانی ٹیم ویسٹ انڈیز کو سان نہ سمجھے Do not underestimate the Pakistani team West Indies	پاکستان لائن سپر لیگ کے دوسرے ایڈیشن میں کھلاڑیوں کی نیلامی کا عمل The process of auctioning players in the second edition of Pakistan Line Super League	اکشے کمار کی نئی ایکشن فلم کے پوسٹر سوشل میڈیا پر جاری The posters of Akshay Kumar's new action film are released on social media	5

Fig. 5. Generated sample for Low-resource language and translation in English using Google Translate.

Table 2
Generated sample for High-resource language.

Original headline	Generated headline before HPPE	Generated headline after HPPE
Man charged over fatal crash	man on trial accused of sex assault	man to face court over bashing
Conlon stands firm on new grid costs	china plans to continue	nurses union slams hospital pay offer
Police cracking down on driver safety	police find missing teen	police hunt for man after stabbing
Chinese taxi drivers strike over fines gas supplies	Chinese Australia and us sign new security asea Taiwan	police investigate Albany death
Harkins told to run as senate candidate	police investigate suspicious death of man	police investigate alleged car crash

size for each generated sample was about 1000 headlines. Further, we have found repeated patterns that occurred consistently. In order to get the top 25 headlines having similar patterns, we have removed POS Tags that were repeating consistently in each headline as shown in Fig. 9. We found some original/generated sequences were having duplication in POS Tags, we removed those patterns too. Once we have done these steps, we compared the pattern of original headlines and GPT2-generated headlines. We kept those GPT-2 generated headlines that were matched with the top 25 POS Tag patterns. 200 out of 1000 were matched with one or many of the top 25 original POS Tags patterns. Therefore, we considered 200 matched GPT-2 generated headlines for further evaluation.

3.5. Turing test

Once we completed the headlines generation process, we mixed it with the original headlines and conducted the Turing test for subjective evaluation. We conducted the Turing test to validate the generated Urdu and English news headlines. The participants, involved in the Turing test were proficient in Urdu and English Language. For the Urdu text evaluation, we selected BS (CS)-II students from Sukkur IBA University, and for the English text, we selected English Faculty staff members from Sukkur IBA University.

In this evaluation, we selected 40 generated Urdu headlines and 40 actual Urdu headlines then we made four groups of participants, namely (A, B, C, and D) and each group had 7 participants thus the total number of participants was 28. Each of the participants was assigned 20

کراچی اردو پوائنٹ تازہ ترین اخبار24اپریل 2019ء قومی کرکٹ ٹیم کے سابق Karachi Urdu Point Latest Newspaper 24 April 2019 Former National Cricket Team	1
واپڈا کے سینئر راونڈر محمد شبیر انٹرنیشنل کرکٹ میں زیادہ مرتبہ WAPDA's senior rounder Mohammad Shabbir has played the most times in international cricket	2
لاہور نمائندہ سپورٹس پی سی بی گورننگ بورڈ کا 52 واں اجلاس اج دوبئ 52nd Meeting of Lahore Rep. Sports PCB Governing Board Aaj Dubai	3
اسٹاف رپورٹ ایڈیلیڈ ئر لینڈ کے گریم سوان کی شاندار نصف سنچر Staff Report Adelaide Ireland's Graeme Swann's brilliant half-century	4
کراچیملک میں تقریبا پانچ ہزار فیکٹریوں کے بجلی کے بحران About 5,000 factories in Karachi are experiencing power outages	5

Fig. 6. Generated Urdu headline sample size for 80 characters and translation in English using Google Translate.

فائل فوٹوسرگودھاچیمبرزکامرس اینڈ انڈسٹری کی ریجنل قائمہ کمیٹی کے چیئرمی File photo Chairman of the Regional Standing Committee of Sargodha Chambers of Commerce and Industry	1
پاکستانی اداکارہ عائزہ خان اور دانش تیمور کی فلم ہولی وڈ کی ریلیز کی تاریخ ک Pakistani actress Ayeza Khan and Danish Taimur's film Hollywood release date	2
بھارتی ٹینس اسٹار ثانیہ مرزا نے نجی مصروفیات کے سبب چھٹی مار کر اپنی زندگی کا Indian tennis star Sania Mirza took a break from her life due to personal engagements	3
ئی ایم ایف کے مطابق پاکستان کے جاری کھاتوں کا توازن بہت بڑھ گیا ہے ئی ایم According to the IMF, Pakistan's current account balance has increased significantly	4
ٹوکیو 92 نیوز ایشین سنوکر ٹیم چیمپئن شپ میں شرکت کے لئے جاپانی ٹیم واہگہ با Tokyo 92 News The Japanese team Wahga Ba to participate in the Asian Snooker Team Championship	5

Fig. 7. Generated Urdu headline sample size for 100 characters and translation in English using Google Translate.

pieces of news (10 generated and 10 actual news) to evaluate if these are generated or actual news. The majority voting scheme was used to finalize the decision of whether the news belongs to generated or actual category. The same process we have done for English headlines.

In order to assess the impact of our proposed post-processing approach, we performed the same experiment on the same level of participants twice: during the first experiment, we mixed generated

شیا کی وزارت داخلہ کے تعاون سے پاکستان کا ایشیا اور ایشین ممالک کے درمیان دو طرفہ ڈیوٹی Pakistan's bilateral duty between Asia and Asian countries with the support of Shea's Ministry of Interior	1
لاہور جسارت نیوز پاکستان کی پہلی انٹرنیشنل کبڈی لیگ اکتوبر میں اسلام باد میں کھیلی جائے گی Lahore Jasrat News Pakistan's first International Kabaddi League will be played in Islamabad in October	2
دبئی این پی قومی کرکٹ ٹیم کے کپتان سرفراز احمدنے کہا ہے کہ پاکستان کی ٹی ٹوئنٹی ٹیم Dubai NP national cricket team captain Sarfaraz Ahmed has said that Pakistan's T20 team	3
سجل علی جلد ہی بہت جلد تینوں فلمی کیریئر کا اغاز کریں گی جس میں وہ خواتین کو اپنی کار Sajal Ali will soon embark on a trio of film careers in which she will give women her car	4
لاہور ڈیلی پاکستان لائن سری لنکا کے خلاف ٹیسٹ سیریز کیلئے قومی ٹیم کا اعلان کر دیا گیا ہے Lahore Daily Pakistan Line has announced the national team for the Test series against Sri Lanka	5

Fig. 8. Generated Urdu headline sample size for 120 characters and translation in English using Google Translate.

news with original news without the involvement of our proposed POS tagging technique which we used as a post-processing step. In the second experiment, we mixed generated news with original news with the involvement of the POS tagging technique proposed in this paper.

4. Results and discussions

In this section, we present our experimental results on low and high-resource languages with and without headline POS pattern extraction (HPPE). The code used for conducting experiments in this paper is publicly available at GitHub.[10]

4.1. Results obtained for low-resource on headlines text

In this setting, we have generated the news headlines from the headlines corpus of the one million dataset as discussed in Section 3.3.1. This corpus contained only headline text, it did not include full news stories. As it can be seen from Fig. 10, the overall results of generated headline news were poor. Semantic and syntax errors were found in generated headlines news. The major reason for the poor performance is that headline size is shorter and GPT-2 required more data to generate quality text (Ko and Li, 2020). Another reason could be GPT-2 is trained on English corpus, it may be difficult for the GPT model to generate text in low-resource language (de Vries and Nissim, 2020). Therefore, to improve the quality of generated headlines, we have unitized the news corpus from one million Urdu dataset as discussed in the next section.

4.2. Results obtained for low-resource language before HPPE on full news story text

In this setting, we generated the headlines from the entire Urdu news dataset. The Turing test results obtained from this setting are shown in Table 3. We have found that 60% of the respondents identified the original news headline as original and 40% of the participants identified it as generated. Similarly, for generated news headlines, 25% predicted as original and 75% predicted as a generated news headline.

In addition to this, we have shown the individual group performance in Fig. 11. For instance, Group A found 80% actual headlines as actual and 20% as generated. Similarly, 80% of generated news headlines were understood as generated news. Moreover, Group B identified all the actual news as actual, and 20% of generated news was predicted as actual. For Group C, we found that 60% of the original news headline was predicted as generated and 40% of generated news as the original one. Likewise, 80% of original headline news was predicted as generated news headlines, and the same ratio was predicted correctly that it was being generated from the model by Group D.

[10] https://github.com/saifhassan/Headline-Generationg-using-POS.git

Table 3
Obtained from human evaluation for Urdu generation before applying HPPE.

	Original news	Generated news
Predicted original news	60%	40%
Predicted generated news	25%	75%

Table 4
Obtained from human evaluation for low-resource language headlines generation after applying HPPE.

	Original news	Generated news
Predicted original news	42%	58%
Predicted generated news	28%	72%

Table 5
Obtained from human evaluation For English headline generation before applying HPPE.

	Original news	Generated news
Predicted original news	42%	58%
Predicted generated news	35%	65%

4.3. Results obtained for low-resource language after HPPE

In this setting, we have applied our proposed headline POS pattern extraction method (HPPE) to fine-tune the headline news generation in order to generate synthetic news. The detail of Turing test results is shown in Table 4. After applying the HPPE, we found significant improvement in the results. We have found that 58% identified the actual headline news as generated and 42% as actual. The generated news was too similar to the actual news that all four groups had difficulty in differentiating the headline news between the actual and generated. The most important is the difference of 3% that is observed in predicted values for generated news. Before our proposed HPPE post-processing approach, 25% of participants incorrectly predicted generated news as actual news, whereas after applying HPPE, this number increased to 28% which means that HPPE post-processing was able to improve generated text quality that additional 3% participants classified generated news as actual news. The individual group performance is shown in Fig. 12.

4.4. Results obtained for high resource language before HPPE

In this setting, we have generated the headlines from the entire English news dataset. The Turing test results obtained from this setting are shown in Table 5. We found that 42% of the people identified the original news as original and 58% of the people identified original news as generated. Similarly, for generated news headlines, 35% predicted as original and 65% predicted as a generated news headline.

In addition to this, we have shown the individual group performance in Fig. 13. For instance, Group B found 70% of actual generated headlines as actual and 30% as generated. Similarly, 20% of generated news headlines were understood as actual news, and the remaining 80% correctly identified as generated headlines. The major reason for predicting the generated headline easily was because our model generated the text incomplete and somehow meaningless as can be seen from Fig. 5. Therefore, to improve the quality of generated headline text, we introduced HPPE in Section 3.4.

headlines	pos_tags	Identical pos_tags
عالمی بینک عسکریت پسندی سے متاثرہ خاندانوں کی معاونت کرے گا The World Bank will support families affected by militancy	ADJ NOUN NOUN NOUN ADP ADJ NOUN ADP NOUN VERB AUX	ADJ NOUN ADP ADJ NOUN ADP NOUN VERB AUX
مالی سال 2020 ریٹرن فائل کرنے والوں کی تعداد میں 23 فیصد کمی 23% decrease in number of FY 2020 return filers	ADJ NOUN NUM NOUN NOUN VERB ADP ADP NOUN ADP NUM NOUN NOUN	ADJ NOUN NUM NOUN VERB ADP NOUN ADP NUM NOUN
جاپان کو سندھ کے خصوصی اقتصادی زون میں سرمایہ کاری کی دعوت Invitation to Japan to invest in Special Economic Zone of Sindh	PROPN ADP PROPN ADP ADJ ADJ NOUN ADP ADJ NOUN ADP NOUN	PROPN ADP PROPN ADP ADJ NOUN ADP ADJ NOUN ADP NOUN
برامدات 767 فیصد بڑھ کر ارب 16 کروڑ ڈالر سے زائد Exports increased by 767% to more than 1.6 billion dollars	NOUN NUM NOUN VERB AUX NUM NUM NUM NOUN ADP ADJ	NOUN NUM NOUN VERB AUX NUM NOUN ADP ADJ
کے الیکٹرک کو اضافی بجلی گیس کی فراہمی کے قانونی تقاضے تعطل کا شکار Legal requirements for supply of additional electricity and gas to K Electric stalled	ADP NOUN ADP ADJ NOUN NOUN ADP NOUN ADP ADJ NOUN NOUN ADP NOUN	ADP NOUN ADP ADJ NOUN ADP NOUN ADP ADJ NOUN ADP NOUN
کھانے پینے کی اشیا کی قیمتیں سال کی بلند ترین سطح پر پہنچ گئیں اقوام متحدہ Food prices hit year's highest, says UN	VERB VERB ADP NOUN ADP NOUN NOUN ADP ADJ ADJ NOUN ADP VERB AUX PROPN PROPN	VERB ADP NOUN ADP NOUN ADP ADJ NOUN ADP VERB AUX PROPN

Fig. 9. Original Urdu headline along with its POS Tag and after removing duplicates consecutively and translation in English using Google Translate.

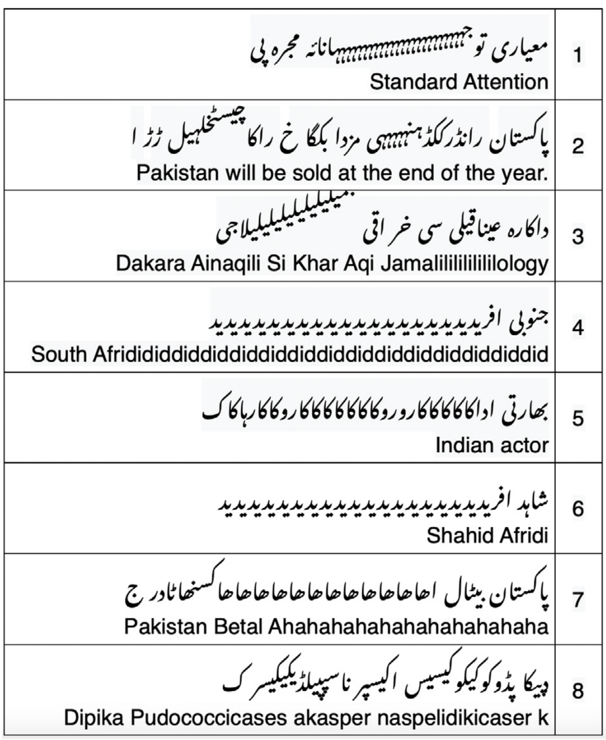

Fig. 10. Low-resources Headline Generation from Headline corpus and translation in English using Google Translate.

4.5. Results obtained for high-resource language after HPPE

In this setting, we have applied the proposed headline POS pattern extraction method to fine-tune the headline news generation in order to generate synthetic news. The results of the Turing test are shown in Table 6. After applying the HPPE, we found significant improvement in the results. We have found that 37% of the people identified the original news as original and 63% of the people identified it as generated. Similarly, for generated news headlines, 82% predicted as original and 18% predicted as generated news headlines. The generated news was too similar to the actual news that all four groups had difficulty differentiating the headline news between the actual and generated. Therefore, there is a decrement of 5% in actual headlines news and an increment of 47% in a generated headline as actual. The individual group performance is shown in Fig. 14

Table 6
Obtained from human evaluation for high-resource language headline generation after applying POST processing.

	Original news	Generated news
Predicted original news	37%	63%
Predicted generated news	82%	18%

4.6. Evaluating generated text

This section explains various machine-centric objective evaluation metrics to evaluate the generated headlines. We compute the quality and similarity of generated headline text to all of the reference/original text as a corpus. We have applied the word-overlap metrics as explained below:

- BLEU: Bilingual Evaluation Understudy: It compares the similarity of the generated text based on n-grams (Papineni et al., 2002). Mathematically, it is defined in Eq. (4):

$$BLEU - N = BP * exp\left(\sum_{n}^{N} w_n log(p_n)\right) \tag{4}$$

  where, N represents the maximum length for the n-gram (in this paper, we have used BLEU-1 and BLEU-2), w represents uniform weighted, and BP shows brevity penalty.
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation: It also compares the similarity of the generated text based on n-gram. But the main difference between ROUGE and BLEU is the former calculates the score on the basis of recall, whereas the latter calculates the F-measure (Lin, 2004).

The results of BLEU and ROUGE scores are presented in Table 7. It can be observed that for both low-resource as well as high-resource languages, the BLEU and ROUGE scores have improved after adding the post-processing step. In BLEU, low-resource language post-processing brought an improvement of about 0.21, whereas, for high-resource, the BLEU improvement is 0.12. On the other hand, ROUGE improvement for low-resource language is 0.017, and for high-resource, it is 0.05. The possible reason for not significant improvement in the ROUGE score may be a number of overlapping words in the generated headline with reference to the original headline. Nonetheless, the generated headlines are so coherent and human-like text that domain experts could not realize that generated headlines were actually generated by GPT-2.

4.7. Ablation study

In this section, we have compared generated synthetic headlines with and without applying the HPPE method because we want to assess

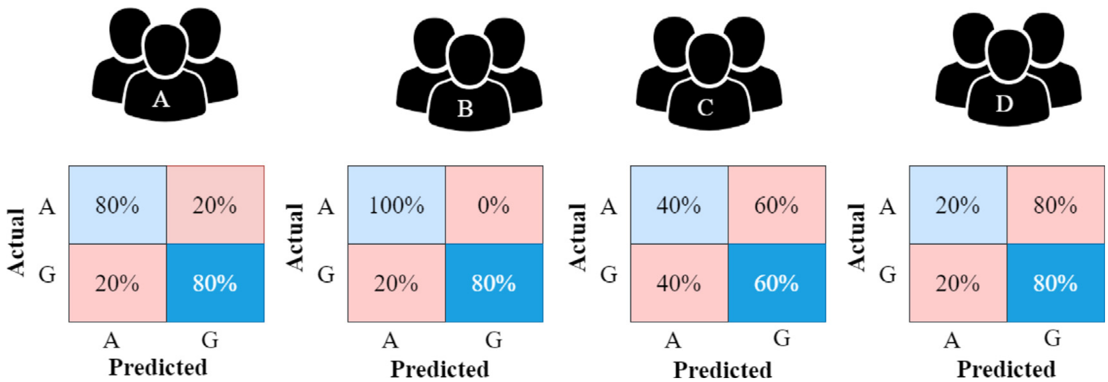

Fig. 11. Individual group performance on low resource language before applying HPPE.

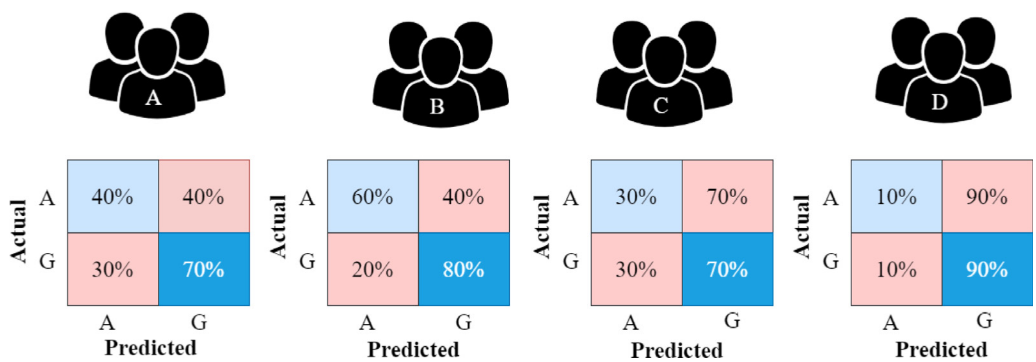

Fig. 12. Individual group performance on low-resource language after applying HPPE.

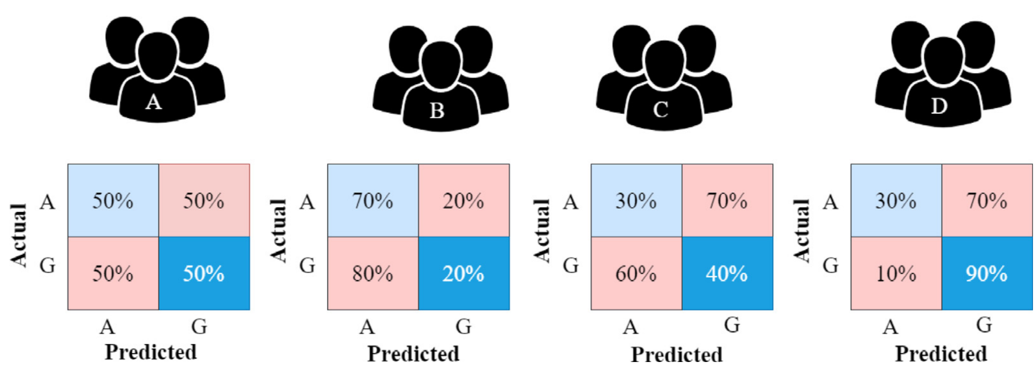

Fig. 13. Individual group performance on High-resource language before applying HPPE.

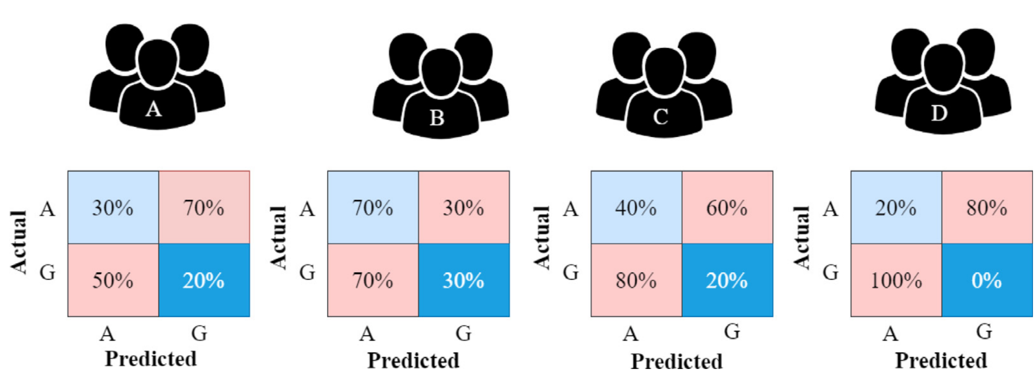

Fig. 14. Individual group performance on High-resource language after applying HPPE.

Table 7
Evaluation metrics scores.

Settings	BLEU	ROUGE
Low-resource language (Before HPPE*)	0.360	0.169
Low-resource language (After HPPE*)	0.570	0.185
High-resource language (Before HPPE*)	0.350	0.105
High-resource language (After HPPE*)	0.470	0.109

*Headlines pos pattern extraction.

Table 8
Top five POS-tag patterns extracted from generated English headlines which passed Turing test.

Antecedent	Consequent	Support	Confidence	Lift
{VERB, NOUN}	{PROPN}	0.3	0.33	1.11
{VERB}	{NOUN, PROPN}	0.3	0.33	1.11
{ADP, PROPN}	{VERB}	0.13	1	1.11
{ADP, NOUN, PROPN}	{VERB}	0.13	1	1.11
{VERB, NOUN}	{ADP, PROPN}	0.13	0.15	1.11

the importance of using the HPPE extraction method in this paper. If we compare the generated headline before (Section 4.2) and after (Section 4.3) applying HPPE for low resource language, we found that after applying HPPE our generated model generated more accurate and similar text to the original once that it is difficult for a person to identify the generated text is original or generated. Therefore, there is an improvement of about 0.21 in the BLEU score and 0.017 in the ROUGE after applying the HPPE.

Similarly, for High resource language, it can be seen in Tables 5 and 6 that there is a decrement of 5% in actual headlines news and an increment of 47% in a generated headline as actual after applying the HPPE. We observe that without HPPE, the model performs poorly because just headline word sequences do not result in contextually important representation.

4.8. Association rules mining for error analysis

To analyze errors in the generated text, we utilized FP-Growth pattern mining to extract frequent POS-tag patterns found in both the generated text that passed the Turing test and those that failed. The top five frequent subsets in English news headlines generated by the model that passed the Turing test are presented in Table 8, while the top five frequent subsets in English news headlines generated by the model that failed in the Turing test are shown in Table 9. Both these subsets were extracted from the generated text after applying our proposed HPPE. Interestingly, all POS patterns presented in Table 8 are subsets of the top ten POS-tag patterns learned from the training data, whereas the patterns in Table 9 are subsets of the bottom five POS-tag patterns learned from the training data. This suggests that if the generated text follows the POS-tag patterns of the training data, it is more likely to pass the Turing test, which is a subjective evaluation due to the improved synthetic appearance of the generated text.

4.9. Key challenges

Some of the potential challenges of the manuscript could be:

- **Selection of POS Patterns:** Choosing appropriate POS patterns to validate the generated headlines can be tricky, as certain patterns may not be applicable to all types of headlines.

Table 9
Top five POS-tag patterns extracted from generated English headlines which failed Turing test.

Antecedent	Consequent	Support	Confidence	Lift
{PROPN, VERB}	{ADP}	0.29	1	1.75
{ADP, VERB}	{PROPN}	0.29	0.5	1.75
{PROPN, NOUN}	{ADP}	0.29	1	1.75
{PROPN, NOUN, VERB}	{ADP}	0.28	1	1.75
{ADP, NOUN, VERB}	{PROPN}	0.28	0.5	1.75

- **Generalizability of Optimal Length:** While the experiments found that 120 characters were the optimal length for generating headlines, this length may not be suitable for all datasets and languages. It is crucial to validate the optimal length on different datasets and languages to ensure that it can produce high-quality and relevant headlines. Additionally, it is important to consider the type of news articles and the target audience when determining the optimal length.
- **Training a Synthetic Headline Generation Model:** Developing a synthetic headline generation model that accurately captures the nuances of both Urdu and English languages can be difficult. It requires extensive pre-processing of the data, fine-tuning of the GPT-2 model, and selecting appropriate hyperparameters.
- **Evaluating the Quality of Generated Headlines:**Evaluating the quality of the generated headlines through human subjective evaluation (Turing test) can be time-consuming and resource-intensive. It also requires selecting an appropriate set of evaluators and avoiding any biases in the evaluation process. Additionally, selecting suitable machine-centric evaluation metrics to compare the generated headlines with the reference headlines can be challenging.
- **Generalization of the Model:** Ensuring the proposed model can generalize to other low-resource and high-resource languages can be a challenge. It requires testing the model on diverse datasets and ensuring that it can produce high-quality headlines that are relevant to a wide range of topics and domains.

4.10. Key advantages

The major key advantages of this work are given below:

- **Novel Contribution:**The manuscript presents a novel approach to synthetic headline generation using GPT-2 for both low-resource language (Urdu) and high-resource language (English). The approach includes extracting POS patterns from original headlines to validate the generated headlines and evaluating the performance through both human subjective evaluation and machine-centric approaches.

- **High Quality of Generated Headlines:** The manuscript claims that the generated headlines using the proposed approach are comparable to the top 25 most common headlines in terms of POS patterns, which indicates that the headlines are of high quality and relevance.

- **Generalizability of the Model:** The manuscript suggests that the proposed approach is generalizable to other low-resource and high-resource languages, which can have significant implications for automating the news headline generation process.

5. Conclusion and future work

Text generation is the ability of AI algorithms to bring creativity among machines. It is an attempt to complete the journey from Artificial Intelligence (AI) to Artificial General Intelligence (AGI). Many recent attempts have encouraged the AI community to finally realize the dream of inducing creativity into algorithms. Recent development like GPTs from OpenAI and BERT from Google has shown promising results in generating creative and synthetic text with just a handful of characters as seed text. Although transformers-based GPT has shown an enormous ability to generate synthetic text, nevertheless it requires abundant data to generate new text into the domain of input data. In case of data scarcity, the generated text suffers from poor quality. This especially is an issue in low-resource languages where labeled data is still scarce. In this paper, we presented a post-processing POS tagging-based step that improves the quality of generated text from GPT in both low-resource as well as high-resource languages. In our proposed model, we first learn frequent POS tag patterns in the original data and train the GPT model to continue generating new instances until they satisfy POS tag patterns found in the original data. We performed different experiments on two publicly available datasets in low- as well as high-resource languages. Furthermore, in order to assess the generated text quality, we performed subjective and objective evaluations. For subjective evaluation, we mixed our model's generated text with the original text and performed the Turing test. The participants of the Turing test were university students and faculty members from the English and Urdu departments. The results of the Turing test indicate that for the English language, there were 47% more respondents who thought our model's generated text is original text after adding our proposed POS tagging step in GPT. For the Urdu language, there were 3% more respondents who thought our model's generated text is original text. Along with subjective evaluation based on the Turing test, we also conducted an objective evaluation of generated text quality through BLEU and ROUGE. Both metrics showed improvement after adding a post-processing step based on POS tagging as proposed in this paper. To the best of our knowledge, this is the first attempt at exploiting GPT for generating Urdu text which is one of the main contributions of this paper. Also, the exploitation of POS tagging as a post-processing step is another significant and generalizable contribution of this paper.

This work can be further extended to other low-resource languages (like Norwegian, Arabic, etc.) where labeling of data is expensive and data resources are scarce. It will be also interesting to see the impact of text generated through the model proposed in this paper for balancing unbalanced datasets and see how it affects the classification accuracy. Although, this work focuses on news headlines generation, however, the post-processing step proposed in this paper can be generalized to other domains like generating story text, music composition, image generation, and other creative aspects of AI.

CRediT authorship contribution statement

Noureen Fatima: Data curation, Methodology, Experimentation, Writing – original draft, Investigation. **Sher Muhammad Daudpota:** Methodology, Supervision, Writing – original draft. **Zenun Kastrati:** Conceptualization, Supervision, Writing – review & editing. **Ali Shariq Imran:** Conceptualization, Supervision, Writing – review & editing. **Saif Hassan:** Data curation, Experimentation, Writing – original draft. **Nouh Sabri Elmitwally:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Department of Computer Science (IDI), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway; and in part by the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) Project NORPART-2021/10502, funded by DIKU, Norway.

We are also thankful to Sukkur IBA University students and faculty members who participated in the Turing Test for the subjective evaluation of generated text before and after applying POS tags.

References

Alfonseca, E., Pighin, D., Garrido, G., 2013. Heady: News headline abstraction through event pattern clustering. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1243–1253.

Ashraf, N., Khan, L., Butt, S., Chang, H.-T., Sidorov, G., Gelbukh, A., 2022. Multi-label emotion classification of Urdu tweets. PeerJ Comput. Sci. 8, e896.

Barros, C., Vicente, M., Lloret, E., 2021. To what extent does content selection affect surface realization in the context of headenhancede generation? Comput. Speech Lang. 67, 101179.

Batra, R., Kastrati, Z., Imran, A.S., Daudpota, S.M., Ghafoor, A., 2021. A large-scale tweet dataset for urdu text sentiment analysis.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L., 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226.

Chandio, B.A., Imran, A.S., Bakhtyar, M., Daudpota, S.M., Baber, J., 2022. Attention-based RU-BiLSTM sentiment analysis model for roman urdu. Appl. Sci. 12 (7), 3641.

Chen, J., Wu, Y., Jia, C., Zheng, H., Huang, G., 2020a. Customizable text generation via conditional text generative adversarial network. Neurocomputing 416, 125–135.

Chen, Y., Yang, C., Liu, Z., Sun, M., et al., 2020b. Reinforced zero-shot cross-lingual neural headline generation. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 2572–2584.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modelin2020g. arXiv preprint arXiv:1412.3555.

de Vries, W., Nissim, M., 2020. As good as new. How to successfully recycle english GPT-2 to make models for other languages. arXiv preprint arXiv:2012.05628.

Fatima, N., Imran, A.S., Kastrati, Z., Daudpota, S.M., Soomro, A., Shaikh, S., 2022a. A systematic literature review on text generation using deep neural network models. IEEE Access.

Fatima, N., Jahangir, R., Mujtaba, G., Akhunzada, A., Shaikh, Z.H., Qureshi, F., 2022b. Multi-modality and feature fusion-based COVID-19 detection through long short-term memory. Comput. Mater. Contin. 4357–4374.

Fujita, K., Watanabe, R., 2018. On implementing an automatic headlin2020e generation for discussion BBS systems—Cases of citizens' deliberations for communities—. IEICE Trans. Inf. Syst. 101 (4), 865–873.

Gavrilov, D., Kalaidin, P., Malykh, V., 2019. Self-attentive model for headlin2020e generation. In: European Conference on Information Retrieval. Springer, pp. 87–93.

Ghafoor, A., Imran, A.S., Daudpota, S.M., Kastrati, Z., Batra, R., Wani, M.A., et al., 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. IEEE Access 9, 124478–124490.

Haidar, M., Rezagholizadeh, M., et al., 2019. Textkd-gan: Text generation using knowledge distillation and generative adversarial networks. In: Canadian Conference on Artificial Intelligence. Springer, pp. 107–118.

Hejazi, H.D., Khamees, A.A., Alshurideh, M., Salloum, S.A., et al., 2021. Arabic text generation: deep learning for poetry synthesis. In: Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021, Vol. 1339, No. 1339. Springer, pp. 104–116.

Hodges, A., 2014. Alan Turing: The Enigma. Princeton University Press.

Hossain, M.R., Hoque, M.M., Siddique, N., Sarker, I.H., 2021. Bengali text document categorization based on very deep convolution neural network. Expert Syst. Appl. 184, 115394.

Hossain, M.R., Hoque, M.M., Siddique, N., Sarker, I.H., 2023. CovTiNet: Covid text identification network using attention-based positional embedding feature fusion. Neural Comput. Appl. 1–25.

Hussain, K., Mughal, N., Ali, I., Hassan, S., Daudpota, S.M., 2021. Urdu news dataset 1M. Mendeley Data 3.

Imran, A.S., Daudpota, S.M., Kastrati, Z., Batra, R., 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 8, 181074–181090.

Imran, A.S., Yang, R., Kastrati, Z., Daudpota, S.M., Shaikh, S., 2022. The impact of synthetic text generation for sentiment analysis using GAN based models. Egypt. Informat. J. 23 (3), 547–557.

Ko, W.-J., Li, J.J., 2020. Assessing discourse relations in language generation from GPT-2. arXiv preprint arXiv:2004.12506.

Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al-Ayyoub, M., Jararweh, Y., 2023. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. Simul. Model. Pract. Theory 102754.

Kulkarni, R., 2018. A Million News Headlines. Harvard Dataverse, http://dx.doi.org/10.7910/DVN/SYBGZL.
Li, P., Yu, J., Chen, J., Guo, B., 2021. HG-news: News headenhancede generation based on a generative pre-training model. IEEE Access 9, 110039–110046.
Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.
Lin, J.C.-W., Shao, Y., Djenouri, Y., Yun, U., 2021. ASRNN: A recurrent neural network with an attention model for sequence labeling. Knowl.-Based Syst. 212, 106548.
Lin, J.C.-W., Shao, Y., Zhang, J., Yun, U., 2020. Enhanced sequence labeling based on latent variable conditional random fields. Neurocomputing 403, 431–440.
Lin, J.C.-W., Shao, Y., Zhou, Y., Pirouz, M., Chen, H.-C., 2019. A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. Eng. Appl. Artif. Intell. 85, 175–181.
Liu, Z., Wang, J., Liang, Z., 2020. Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05. pp. 8425–8432.
Lv, Z., Li, J., Dong, C., Li, H., Xu, Z., 2021. Deep learning in the COVID-19 epidemic: A deep model for urban traffic revitalization index. Data Knowl. Eng. 135, 101912.
Mishra, R., Zhang, S., 2021. POSHAN: Cardinal POS pattern guided attention for news headline incongruence. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 1294–1303.
Mosallanezhad, A., Shu, K., Liu, H., 2020. Topic-preserving synthetic news generation: An adversarial deep reinforcement learning approach. arXiv preprint arXiv:2010.16324.
Mosallanezhad, A., Shu, K., Liu, H., 2021. Generating topic-preserving synthetic news. In: 2021 IEEE International Conference on Big Data. Big Data, pp. 490–499. http://dx.doi.org/10.1109/BigData52589.2021.9671623.
Nishi, Y., Suge, A., Takahashi, H., 2021. Construction of a news article evaluation model utilizing high-frequency data and a large-scale language generation model. SN Bus. Econ. 1 (8), 1–18.
O'Shea, K., Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
Pal, S., Chang, M., Iriarte, M.F., 2022. Summary generation using natural language processing techniques and cosine similarity. In: Intelligent Systems Design and Applications: 21st International Conference on Intelligent Systems Design and Applications (ISDA 2021) Held During December 13–15, 2021. Springer, pp. 508–517.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.
Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D., 2020. Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.
Rumelhart, D.E., Widrow, B., Lehr, M.A., 1994. The basic ideas in neural networks. Commun. ACM 37 (3), 87–93.
Schmidhuber, J., 2015. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. arXiv preprint arXiv:1511.09249.
Shaikh, S., Daudpota, S.M., Imran, A.S., Kastrati, Z., 2021. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Appl. Sci. 11 (2), 869.
Shao, Y., Lin, J.C.-W., Srivastava, G., Jolfaei, A., Guo, D., Hu, Y., 2021. Self-attention-based conditional random fields latent variables model for sequence labeling. Pattern Recognit. Lett. 145, 157–164.
Shavrina, T., Shliazhko, O., 2021. Using generative pretrained transformer-3 models for Russian news clustering and title generation tasks.
Shen, S.-q., Chen, Y., Yang, C., Liu, Z.-y., Sun, M.-s., et al., 2018. Zero-shot cross-lin2020gual neural headlin2020e generation. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12), 2319–2327.
Shen, S.-Q., Lin, Y.-K., Tu, C.-C., Zhao, Y., Liu, Z.-Y., Sun, M.-S., et al., 2017. Recent advances on neural headline generation. J. Comput. Sci. Tech. 32 (4), 768–784.
Shenassa, M.E., Minaei-Bidgoli, B., 2022. ElmNet: a benchmark dataset for generating headlin2020es from Persian papers. Multimedia Tools Appl. 81 (2), 1853–1866.
Singh, A., Josan, G.S., 2022. Apply paraphrase generation for finding and ranking similar news headlines in Punjabi language. J. Sci. Res. 66 (1).
Singh, R.K., Khetarpaul, S., Gorantla, R., Allada, S.G., 2021. SHEG: summarization and headenhancede generation of news articles using deep learning. Neural Comput. Appl. 33 (8), 3251–3265.
Talafha, S., Rekabdar, B., 2021. Poetry generation model via deep learning incorporating extended phonetic and semantic embeddings. In: 2021 IEEE 15th International Conference on Semantic Computing. ICSC, IEEE, pp. 48–55.
Tseng, Y.-C., Yang, M.-H., Fan, Y.-C., Peng, W.-C., Hung, C.-C., 2022. Template-based headenhancede generator for multiple documents. IEEE Access.
Turing, A.M., 2009. Computing machinery and intelligence. In: Parsing the Turing Test. Springer, pp. 23–65.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, N., Issa, R.R., 2020. Natural language generation from building information models for intelligent NLP-based information extraction. In: EG-ICE 2020 Workshop on Intelligent Computing in Engineering. Universitätsverlag der TU Berlin2020, Berlin2020, pp. 275–284.
Wang, Z., Xu, L., Liu, Z., Sun, M., et al., 2020. Topic-sensitive neural headlin2020enhancede generation. Sci. China Inf. Sci. 63 (8), 1–16.
Xie, J., Wang, X., Wang, X., Pang, G., Qin, X., 2019. An eye-tracking attention based model for abstractive text headlin2020e. Cogn. Syst. Res. 58, 253–264.
Xu, Z., Lv, Z., Li, J., Sun, H., Sheng, Z., 2022. A novel perspective on travel demand prediction considering natural environmental and socioeconomic factors. IEEE Intell. Transp. Syst. Mag. 15 (1), 136–159.
Zhu, Q., Luo, J., 2023. Generative transformers for design concept generation. J. Comput. Inf. Sci. Eng. 23 (4), 041003.