Automated News Headline Generation Using Natural Language Processing Techniques

Grace Li dl4610@nyu.edu
Nick Yi iy2046@nyu.edu
Golam Raiyan gr2257@nyu.edu

CSCI-UA 0469: Natural Language Processing

Abstract

News Headlines, with their cultural significance and intricate linguistic structure, present a challenging domain for Natural Language Processing (NLP). In this paper, we present an automated news headline generator that utilizes advanced NLP techniques to address the need for fast, coherent, and contextually relevant news article headlines. Utilizing the functionality of the SpaCy library, our approach integrates Part-of-Speech (POS) tagging and Named Entity Recognition (NER) to extract key phrases and entities for enhanced semantic understanding of news content. We employ a TF-IDF model to emphasize relevant terms and use a cosine similarity metric to assess the validity of headlines to the original articles' context. This approach not only automatically generates headlines, but also ensures that they are meaningful and closely reflect the core themes of the article. Preliminary results from both quantitative measurements and human evaluation show that our model is highly effective in generating concise and informative headlines, demonstrating its potential to support journalists and content creators in streamlining their workflow.

1 Introduction and Motivation

In the rapidly evolving digital news landscape, the ability to swiftly generate engaging and informative headlines is more crucial than ever. Headlines serve not only to capture attention but also to succinctly convey the essence of news articles—influencing how information is perceived and consumed. As the volume of online content grows and the pace of news dissemination accelerates, the automation of headline generation becomes essential, yet remains fraught with challenges. Traditional automated systems often fail to capture the semantic depth and stylistic nuances necessary for creating impactful headlines, relying instead on simplistic extraction of key phrases or sentences. This approach frequently falls short in representing the article's full context or engaging a diverse audience effectively.

Our research addresses these limitations by harnessing advanced NLP techniques—including POS tagging and NER through the comprehensive linguistic capabilities of the SpaCy library. This methodology enhances the semantic analysis of text, enabling the generation of more nuanced and contextually relevant headlines. Additionally, by integrating TF-IDF statistics and applying cosine similarity metrics, our model not only automates but also ensures that the generated headlines maintain a high degree of relevance and engagement potential.

This paper presents an advanced algorithmic approach that uses language analysis to enhance computational journalism. By evaluating the performance of our model with human judgment and quantitative metrics, we provide a comprehensive evaluation of its effectiveness. The expected benefits of this project to news organizations and content creators are huge, providing a flexible tool that can adapt to a variety of news styles and simplify the content creation process. Ultimately, our system aims to support creators in expressing the central idea and style of their articles more efficiently and accurately.

2 Architecture and Methodology

2.1 POS Tagging

Part-of-Speech (POS) tagging is a foundational component of our headline generation system. Utilizing the SpaCy library, this process involves analyzing the words in a news article to identify their corresponding parts of speech, such as nouns, verbs, adjectives, etc. This linguistic analysis helps in structuring and understanding the grammatical composition of the text, which is crucial for accurately extracting key phrases and terms that are essential for headline generation.

2.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) is employed to identify and classify the key entities within the text into predefined categories like person names, organizations, locations, and other specifics. By using SpaCy's robust NER capabilities, our system can highlight significant entities that should be featured in the headlines, ensuring that these critical elements are not overlooked and that the headlines reflect the primary focus of the articles.

2.3 TF-IDF Calculation

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate how important a word is to a document within a corpus. In our system, TF-IDF calculation helps in determining the significance of each word in an article relative to a collection of news articles. By assigning higher weights to terms that are frequent in a specific document but not common across other documents, our model can identify unique keywords that are potential candidates for inclusion in the headlines. TF measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more often in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalization.

$$TF(t) = \left(\frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}\right)$$

IDF measures how important a term is. While computing TF, all terms are considered equally important. However, certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scaling up the rare ones, by computing the following.

$$IDF(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}\right)$$

$$TFIDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

where t is the term, d is the document, and D is the corpus.

2.4 Headline Generation

The title generation module synthesizes the information processed by the previous components to create concise and informative titles. This process strategically combines key phrases, important entities, and important terms identified through POS tagging, NER, and TF-IDF analysis. The algorithm aims to ensure that the generated title not only reflects the content of the article, but is also engaging and coherent. Each of these components plays a vital role to guarantee that the titles generated by our system are of high-quality and that they meet professional journalistic standards.

3 Experiments

3.1 Dataset

The dataset utilized for this study is sourced from Kaggle, including 6,256 news articles. Each article in the dataset is characterized by several attributes, including the Article Title, Article Text, and Article Label. This extensive collection features a variety of news stories curated from different news agencies, providing a broad and representative sample of journalistic content.

The Article Text in the dataset serves a critical role in our evaluation process, as they allow us to compare the system-generated headlines against the texts. This comparison helps assess the effectiveness and relevance of our headline generation algorithm. Furthermore, the Article Text is used as the primary input for our NLP pipeline, where it undergoes Named Entity Recognition (NER), tokenization, and Part-of-Speech (POS) tagging. These processes are essential for extracting the key linguistic elements that inform the headline generation process.

3.2 Data Preprocessing

The preprocessing of the dataset is a critical step in preparing the raw text data for effective analysis and processing by our headline generation system. Here's an overview of the preprocessing steps implemented: The first step involves cleaning the text to remove any artifacts that could interfere with the linguistic analysis. This includes stripping out special characters, and extra whitespace. Such elements are common in scraped web data and can lead to inaccuracies in text processing if not removed. Once the text is cleaned, it is tokenized. Tokenization is the process of splitting the text

into individual elements or tokens. In the context of our project, tokens are typically words, but they can also include punctuation and other elements depending on the granularity required. Tokenization is fundamental because it transforms the raw text into a structured form that can be analyzed at the word or token level.

3.3 Experimental Process

For the purposes of this project, we did not employ a traditional training set due to the nature of our headline generation model, which focuses on linguistic rule-based techniques rather than predictive modeling. Instead, we divided our dataset into two distinct sets:

Development Set: Comprising 50 news articles, this set was used during the initial phase of our project to develop and fine-tune our headline generation system. The development set allowed us to iteratively test and refine our algorithms, ensuring that the system was accurately processing the input text and generating coherent headlines.

Test Set: Consisting of another 50 news articles, the test set was utilized to rigorously evaluate the performance of our headline generation system. This set acted as a new, unseen dataset to simulate real-world conditions where the effectiveness of the generated headlines could be judged.

This experimental configuration, with distinct development and test sets, ensures that our evaluations are robust and that our system is capable of performing consistently across different samples of data.

4 Evaluation

4.1 Quantitative Evaluation

Our quantitative evaluation primarily focuses on measuring the relevance of the generated headlines to the corresponding article texts using cosine similarity. This metric provides a numerical indicator of how closely the semantic content of the generated headlines matches the original articles.

Cosine Similarity

Cosine Similarity Calculation: We convert the text of the original articles and the generated titles into vector form using the TF-IDF technique. This conversion allows us to quantify the importance of words based on their frequency in a specific document relative to their frequency across all documents in our dataset. Once the TF-IDF vectors are constructed for the original texts and generated headlines, cosine similarity scores are calculated to determine the degree of alignment between each pair.

Cosine similarity scores range from 0 to 1, where 1 indicates perfect similarity. These scores are analyzed to evaluate the effectiveness of the title generation process in the development and test sets. We calculate the average cosine similarity across all articles to get a common performance measure. A high average score indicates that our titles closely reflect the substantive content of the articles, indicating a successful title generation strategy.

4.2 Human Evaluation

Human evaluation forms a critical component of our headline generation system's assessment, offering essential insights into the coherence, relevance, and impact of the generated headlines from a user perspective. This evaluation process is particularly vital as it gauges the human readability and appeal of the headlines—factors that are often not fully captured by quantitative metrics. We implemented a two-stage evaluation process to thoroughly assess the generated headlines.

Initial Internal Evaluation

Conducted by our development team of three members, all fluent in English and adept at comprehending news articles, the initial phase of evaluation focused on the early batch of generated headlines. The team independently assessed the first 10 generated headlines for their syntactic correctness, semantic coherence, and overall engagement, rating each on a scale of 1 to 5. This internal review served to pinpoint immediate areas for improvement and provided an early indication of the model's performance, setting the stage for more extensive testing.

Ranking

Evaluation Criteria	The Specification	Score
Syntactic Relation	Correct or not	1-5
Semantics	Fluency or not	1-5
Semantics	Content Relevance or not	1-5
Overall & Engagement	Overall	1-5

Table 1: Human Evaluation Criteria.

Syntactic Relations: Headlines were evaluated for grammatical accuracy to ensure they adhered to standard language conventions.
Semantics: Evaluators assessed whether the headlines were meaningful and semantically aligned with the articles, reflecting the main themes accurately.
Overall & Engagement: This criterion measured the headlines' ability to attract attention and engage readers, an essential aspect of headline effectiveness.

External Evaluation

Following refinements based on feedback from our initial internal review, we conducted a more formal second evaluation with ten native English speakers. In this phase, participants were presented with a set of 10 articles, each accompanied by both its original headline and one generated by our model, arranged in a double-blind format to eliminate any bias toward human-written or AI-generated content. Evaluators were tasked with selecting the headline that they felt best matched the article's content, unaware of which headlines were produced by our system. The selection rate of AI-generated headlines by these participants served as a direct metric for assessing our model's capability to create human-like and contextually appropriate headlines.

5 Evaluation Analysis

5.1 Quantitative Results

To evaluate the performance of our headline generation system, we conducted a quantitative analysis on a testing corpus of 50 news articles. Our evaluation relied on manual review to classify generated headlines into four categories: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN); TPs are cases where the generated headline meets or exceeds the similarity threshold and is rated as accurate in the manual evaluation, FPs occur when the headline meets the threshold but is rated poorly, FNs arise when the headline does not meet the threshold but is rated as high-quality during manual review, and TNs represent cases where the headline neither meets the threshold nor is highly rated. We determined that a cosine similarity threshold of 0.3 was the most appropriate for our evaluation. From the 50 articles evaluated, the system produced 29 TPs, 4 FPs, 15 FNs, and 2 TNs. These results were used to calculate the system's precision, recall, and F-measure, as outlined in the equations below:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}, \text{F-measure} = 2\times\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision was computed as 0.8788; **Recall** was slightly lower at 0.6591; and the **F-measure** was calculated as 0.7532. These results demonstrate that the system performs well in generating relevant headlines.

5.2 Human Evaluation Analysis

Our two-stage human evaluation of the title generation project provided comprehensive insights into the immediate impact and iterative improvements of our system. The following is an in-depth analysis of how this evaluation strategy helped improve the title generator and the implications of the results for its current and future features.

5.2.1 Internal Evaluation Analysis

An initial internal evaluation by our development team helped identify and address immediate deficiencies in the algorithm. Feedback on syntactic correctness, semantic coherence, and engagement enabled us to make targeted improvements that would improve the linguistic accuracy and relevance of the generated headlines. This early tuning phase was critical as it laid a solid foundation for subsequent evaluations, ensuring that the headline generator adhered more closely to language norms and reader expectations from the outset.
The results of the first phase directly influenced enhancements to the headline generation algorithm, demonstrating the panelists' ability to adjust and evolve; this adaptability is a key strength of our development process.

5.2.2 External Evaluation Analysis

In the external evaluation, involving ten native English speakers, a double-blind setting was used to ensure that the feedback was unbiased and based solely on the raters' perception of headline quality. The frequency with which AI-generated headlines were selected compared to the original headlines provided a clear quantitative measure of the model's success in replicating professional headline standards. This is critical for evaluating the effectiveness of the system from a broader, more public perspective.

The fact that multiple AI-generated headlines were selected over human-written headlines suggests that the model can not only mimic, but sometimes even surpass human performance in terms of relevance and appeal. This result speaks volumes about the potential application of the system in real-world settings, where it can help journalists and content creators by providing high-quality headline suggestions.

5.2.3 Overall

Our News Headline Generation underwent a comprehensive human evaluation in two different phases, which not only verified the quality effectiveness of the generated titles, but also highlighted the system's ability to generate content that resonates deeply with human readers. The development process was enriched by incorporating detailed feedback from internal team members and external evaluators, significantly improving the linguistic quality and reader engagement of titles. This two-stage evaluation not only confirmed the current performance of the title generator, but also paved the way for future improvements. As we advance, incorporating more linguistic features and deeper semantic analysis can unlock even more powerful capabilities.

6 Conclusion

This study explored the development of an automatic news headline generation system using advanced natural language processing techniques. Although the system has demonstrated good capabilities in generating linguistically accurate and contextually relevant headlines, significant challenges and areas for improvement remain.

Despite highlighting successes through quantitative and human evaluations, the system also generates headlines that sometimes lack depth or fail to fully capture the content of the article. These deficiencies often manifest as headlines that fail to fully reflect the news story. Such inconsistencies indicate a need to enhance the model's understanding and processing capabilities, and the system could benefit from enhanced semantic understanding.

Looking forward, the future improvements could be to implement an automatic evaluation system that can assess the overall topic and structural coherence of the entire news headline—just as humans do—in order to better incorporate semantic depth and narrative style into the headline generation process. This would involve improving our model's ability to integrate and consider the content of all generated lines when generating each news headline, to ensure that the output is cohesive and contextually relevant. In addition, updating the system to intake an expansive training dataset—that covers a wide range of news types and styles—may help improve the system's ability to handle a wider range of news content and achieve the level of sophistication required for high-quality news reporting.

In conclusion, our project not only lays a solid foundation for automatic news headline generation, but also demonstrates its ability to solve problems efficiently as a technical model. Through continued research and development, we are optimistic about the future capabilities of this tool in the fields of computational creativity and computer journalism.

References

Cai, Pengshan, et al. Generating User-Engaging News Headlines, aclanthology.org/2023.acl-long.183.pdf. Accessed 10 Nov. 2024.

Fatima, Noureen, et al. "Improving News Headline Text Generation Quality through Frequent Pos-Tag Patterns Analysis." Engineering Applications of Artificial Intelligence, Pergamon, 17 July 2023, www.sciencedirect.com/science/article/pii/S0952197623009028?via%3Dihub.

Laban, Philippe, et al. "News Headline Grouping as a Challenging NLU Task." ACL Anthology, aclanthology.org/2021.naacl-main.255/. Accessed 10 Nov. 2024.

Mathur, Prashant, et al. "Generating Titles for Millions of Browse Pages on an E-Commerce Site." ACL Anthology, aclanthology.org/W17-3525/. Accessed 10 Nov. 2024.

Mihalcea, Rada, and Paul Tarau. TextRank: Bringing Order into Texts, aclanthology.org/W04-3252.pdf. Accessed 10 Nov. 2024.

News Headline Generator Dataset

GitHub Repository for News Headline Generator Code