Linguistic Feature-Based Clickbait Detection in Taiwanese News Headlines

Chiung-Wen Chang[1]*, Ching-Han Huang[2],
[1]Graduate Institute of Linguistics, Phonetics and Psycholinguistics Laboratory, National Chengchi University, Taipei, Taiwan
[2]Graduate Institute of Linguistics, Knowledge Understanding and Language Processing Laboratory, National Chengchi University, Taipei, Taiwan
Correspondence: cwchang000@gmail.com

Abstract

This study investigates the use of linguistic features to enhance clickbait detection in traditional Chinese news headlines from Taiwanese media. While clickbait detection has been extensively explored in English, research on Chinese—especially in the context of Taiwanese media—remains sparse. Existing studies often focus on simplified Chinese from Chinese media, which may not accurately reflect the cultural and linguistic nuances of Taiwanese news. This research applies linguistic features, such as forward-reference, listicle formats, and suspenseful or exaggerated language, to improve clickbait detection using neural network models. The study's dataset consists of real online news headlines in Taiwan, and models including RNN, LSTM, GRU, and their bidirectional variants were employed in the analysis. The Bi-GRU model performed best, with linguistic features further improving accuracy to 0.75. This study contributes to the field by utilizing deep learning on a traditional Chinese dataset and demonstrates the value of linguistic features in enhancing model accuracy.

1 Introduction

The title of an article plays a crucial role in summarizing its content and enabling readers to quickly assess its relevance (Scott, 2021). However, in an era of information overload, people have limited attention to spare for articles. As a result, certain media employ manipulated headlines, commonly known as clickbait, to lure readers into clicking on their content. Subsequently, readers may realize that the actual article content does not align with their initial expectations. Clickbait refers to " content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page " (Chen et al., 2015). This technique creates an " information gap " and conceals the core essence of the article by presenting events in an ambiguous manner to entice readers' clicks (Loewenstein, 1994).

It is important to distinguish clickbait from fake news, as the key distinction lies not in the authenticity of the content, but in the gap between the headline and the article content. These intriguing statements, lacking clear explanations, entice readers' curiosity and create a curiosity gap (Loewenstein, 1994; Scott, 2021). The readers do not know what exactly happened until they click on the article. It is a trap that many people have fallen into, and several studies have pointed out that clickbait headlines make people feel cheated and uncomfortable (Beleslin et al., 2017; Chen et al., 2015; Shinkhede, 2019; Jung et al., 2022). However, distinguishing clickbait titles from conventional ones may be possible due to their distinct writing style. Blom and Hansen (2015) argue that clickbait employs stylistic and narrative techniques as diversions, while propose four presentation variables for clickbait: incomplete information, appealing expressions, repetition and serialization, and exaggeration.

Previous research suggests that linguistic clues can be used to identify these writing differences. Clickbait often utilizes the forward-reference technique to imply the existence of highly relevant information without actually providing it. Therefore, unresolved pronouns including demonstrative pronouns, personal pronouns, deictic words, and deixis, commonly appear in clickbait titles. (Bazaco et al., 2019; Blom and Hansen, 2015; Shinkhede, 2019). Additionally, clickbait employs the listicle format to attract readers (Vijgen et al., 2014). Listicle headlines present articles in a list format, indicating the number of items and the list's theme in the title. However, readers cannot only understand the actual content of the list from the title and they must click to access the complete list. (Bazaco et al., 2019). Suspenseful words and exaggerated words are also common characteristics of clickbait (Lun, 2021). Suspenseful words, such as "reveal," "uncover," and "expose," create an anticipation of secrets being unveiled. These terms are strategically used in clickbait headlines to entice readers by promising to solve mysteries or disclose complete information. Exaggerated words employ imaginative language to captivate readers' attention (Bazaco et al., 2019). To sum up, the utilization of linguistic cues holds great potential in facilitating the detection of clickbait and providing individuals in avoiding its associated pitfalls.

2 Related work

Early clickbait detection tasks involved binary classification using traditional supervised models with feature extraction. In early research, Potthast et al. (2016) constucted an English clickbait corpus using Twitter tweets from the top 20 most prolific publishers, containing well-known English newspapers publishers such as BBC News. Three annotators categorized the data into clickbait and non-clickbait categories. Features such as teaser messages, linked web page, and meta information are extracted for model training. The performance of three machine learning algorithms: Logistic Regression, Naive Bayes, and Random Forest was compared. Meanwhile, Chakraborty et al. (2016) collected non-clickbait data from Wikinews and clickbait data from other news media to develop a browser add-on for detection. Features based on sentence structure, word patterns, clickbait language, and n-gram were adopted for model training, employing Decision Tree (DT), Random Forests (RFs), and Support Vector Machine (SVM) as learning algorithms. With the development of neural networks, more clickbait detection tasks have been conducted using deep learning models. Chawda et al. (2019) employed neural network algorithms, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), for the clickbait detection tasks. Additionally, they incorporated a Recurrent Convolutional Neural Network to capture contextual information. Their findings suggested that deep learning algorithm models outperformed traditional supervised algorithm model, such as SVM.

While the majority of studies on clickbait detection have concentrated on English texts, there has been relatively little exploration into Chinese texts. Liu et al. (2021) addressed this gap by constructing a clickbait dataset from WeChat, a Chinese social media platform, focusing on news headlines. They labeled the data into three categories—non-clickbait, general clickbait, and malicious-clickbait, which included vulgar or pornographic titles—using a three-person majority vote. Subsequently, Liu et al. (2022) further expanded on this by exploring the extraction of semantic and syntactic information for training, with both traditional and deep learning algorithms, such as Bidirectional Encoder Representation from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks, showing superior performance.

However, even among the limited studies on Chinese texts, the focus has predominantly been on simplified Chinese used in mainland Chinese online media. This approach may not accurately capture the nuances of clickbait in traditional Chinese news due to cultural and linguistic differences. Therefore, this study aims to address these gaps by conducting clickbait detection on traditional Chinese news headlines from Taiwanese media. The objective is to investigate if linguistic features identified in previous studies on English texts can enhance the automatic classification of clickbait in traditional Chinese contexts.

3 Methodology

3.1 Dataset

A total of 1010 news headlines were collected from Nownews, a Taiwanese news media known for providing the latest news with the fastest updating rate. Three annotators were taught the principle of clickbait, and a majority vote was conducted to classify the data into two categories: clickbait and non-clickbait. The annotation results revealed an imbalanced dataset, consisting of 275 clickbait

Feature Category	Description	Examples
Forward-reference	Demonstrative pronouns, personal pronouns	他/她 (he/she), 這 (this), 那 (that)
Listicle	Numbers	一 (one), 二 (two), 三 (three)
Suspenseful words	The words revealing the secret to create suspense	疑 (doubt), 曝 (expose), 露 (reveal), 公開 (unveil)
Exaggerated words	Emotional punctuations and words	！ (exclamation mark), ？ (question mark), 驚 (shock), 轟 (boom)

Table 1: Categories of handcrafted linguistic features

and 735 non-clickbait headlines. Due to the limited data size, oversampling was not performed to avoid overfitting. Instead, random undersampling was applied, resulting in a total of 550 news headlines, evenly distributed with 275 in each category. Subsequently, the dataset was split into an 80:20 ratio, where 80% of the data was used for training, and 20% for testing. A random seed was set for reproducibility.

3.2 Embedding

The data underwent preprocessing, retaining only the characters, and then tokenization was performed. The tokenized words were converted into word vectors capable of capturing word semantics, which served as text features for model training. Subsequently, CKIP Glove, a pre-trained embedding, was employed. CKIP Glove was a word embedding trained on the Chinese GigaWord Corpus and the Academia Sinica Balanced Corpus of Modern Chinese. It consists of 300-dimensional word vectors (Chen and Ma, 2017, 2018; Fan et al., 2019).

3.3 Feature Extraction

Table 1 presents the categories of handcrafted linguistic features, including forward-reference, listicle, suspenseful words, and exaggerated words.

Forward-reference such as, personal pronouns “他/她” (he/she), and demonstrative pronouns “這” (this) and “那” (that), introduce a curiosity gap, enticing audiences to click on the associated links, while suspenseful and exaggerated words are frequently employed to make sense of drama and attract readers ’ curiosity(Jung et al., 2022; Scott, 2021).

3.4 Training

Deep learning algorithms, Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), have been commonly employed in clickbait detection tasks (Chawda et al., 2019; Liu et al., 2021). In this study we adopted these neural network algorithms: (1)RNN (2)LSTM (3)GRU, as well as their bidirectional counterparts: (1)Bi-RNN (2)Bi-LSTM (3)Bi-GRU. The baseline models were using text vectors and the pre-trained embedding, with the inclusion of optimizers. After training, the baseline models were compared to each other. The model had the best performance was selected for further training, incorporating handcrafted linguistics features.

3.5 Evaluation

Following the training phase, the baseline models were evaluated based on accuracy and F1-score to determine the best-performing model one for the second phase of training, which included handcrafted linguistic features. After the second phase of training, the model was evaluated in terms of precision and recall for further error analysis.

4 Results

Table 2 presents the performance of various deep learning models in clickbait detection. The models were evaluated based on their accuracy and F1 score. Among these six baseline models, the Bi-GRU model with Glove embeddings demonstrated superior performance, achieving an accuracy of 0.74 and an F1 score of 0.73. To further enhance its performance, the Bi-GRU baseline model and was augmented with hand-crafted linguistic features. The resulting model, referred to as Bi-GRU with Glove embeddings and hand-crafted linguistic features, achieved the highest perfor-

Model	Accuracy	F1 score
LSTM + Glove (pretrained embedding)	0.70	0.70
GRU + Glove	0.69	0.70
RNN + Glove	0.67	0.70
Bi-LSTM + Glove	0.70	0.71
Bi-GRU + Glove	**0.74**	**0.73**
Bi-RNN + Glove	0.67	0.65
Bi-GRU + Glove + hand-crafted linguistic features	**0.75**	**0.74**

Table 2: Performance comparison of different models using Glove embeddings.

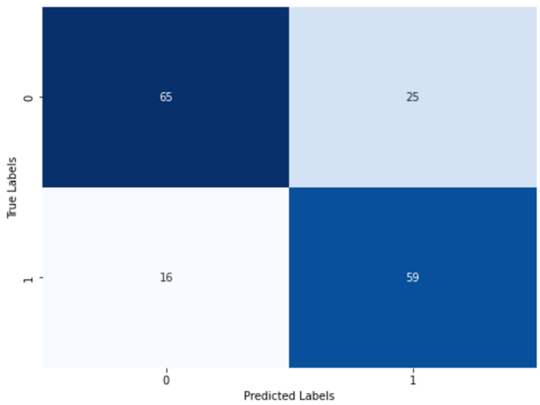

Figure 1: Confusion Matrix of Bi-GRU model with Glove embeddings and hand-crafted linguistic features.

mance with an accuracy of 0.75 and an F1 score of 0.74. This model will be further analyzed in subsequent steps. Figure 1 illustrates the confusion matrix of the Bi-GRU model with Glove embeddings and hand-crafted linguistic features.

Based on the confusion matrix, the precision of the model is calculated as 0.70, indicating that among the predicted positive clickbait instances, 70.2% were actually clickbait. The recall of the model is calculated as 0.79, indicating that the model identified 78.6% of the actual clickbait instances. The F1-score, which combines precision and recall, is calculated as 0.74. Overall, the results demonstrate promising capabilities of the Bi-GRU model with Glove embeddings and hand-crafted linguistic features in clickbait detection.

5 Discussion

The model exhibits a lower Type II error rate, implying fewer false negatives. This suggests a higher recall, indicating that most of the actual clickbait headlines are successfully detected. Conversely, the model demonstrates a higher Type I error rate, resulting in more false positives. Only 70.2% of the headlines predicted as clickbait were actually clickbait. This training outcome suggests that the model may exhibit overgeneralization, classifying more non-clickbait headlines as clickbait, thereby mistakenly identifying some non-clickbait instances.

Upon further examination of the model's prediction errors, particularly within the Type I error category, it is evident that the model often misclassifies non-clickbait headlines containing emotive punctuation marks such as exclamation and question marks as clickbait. This finding aligns with an observed trend where certain non-clickbait headlines, which lack a hook and present content directly related to the headline, are still misclassified as clickbait due to the presence of these exaggerated punctuations. This suggests that emotive punctuations, while often present in clickbait, are also common in non-clickbait Chinese news headlines, reflecting a broader stylistic convention in Chinese journalism that the model has not yet differentiated effectively. Optimization focusing on reducing reliance on emotive punctuation for classification may effectively decrease false positives, leading to a substantial improvement in precision and overall recognition capability.

Additionally, in the Type II error category, it is observed that certain clickbait headlines employ provocative verbs to describe an event without explicitly revealing its nature. However, the model misclassifies them as non-clickbait. This could be attributed to the rich vocabulary and creative phrasing often employed in Chinese news headlines. The model's misclassification of such headlines may indicate a gap in its training corpus, where it may not have learned sufficient vocabulary or contextual nuances. To address this, expanding the lexicon used in hand-crafted linguistic features by collecting diverse vocabulary from news-related corpora could potentially reduce false negatives and increase the model's recall.

To sum up, the Bi-GRU model with Glove embeddings and hand-crafted linguistic features exhibits promising performance in clickbait detection. However, optimization strategies that address both the overgeneralization towards emotive punctuation in non-clickbait headlines and the vocabulary gaps that lead to missed clickbait headlines could significantly enhance precision and recall, leading to improved overall model performance.

6 Conclusion

In conclusion, we conducted clickbait detection training using deep learning models on news headlines from Taiwanese media, with the Bi-GRU model demonstrating the best performance among the neural networks tested. While the inclusion of handcrafted linguistic features improved the model's performance, several limitations emerged. The linguistic features employed in previous studies were primarily based on English data, which presents challenges when applied to Chinese. For instance, Chinese characters can carry multiple meanings depending on the context, unlike English, which typically uses fixed vocabulary for specific meanings. Additionally, a single character may represent various meanings in Chinese, leading to potential confusion when these characters are combined. This issue is further compounded in Chinese news headlines, which often abbreviate words by omitting one character from a two-character term, a phenomenon unique to the language. Such abbreviations can deepen the challenges of text comprehension for models. Similarly, different words in Chinese may convey similar meanings, adding another layer of complexity to feature extraction.

Our future work will focus on refining feature extraction methods, including developing specialized tokenizers and expanding the training dataset. We will also explore the impact of exaggerated words and emotive punctuation on clickbait detection and investigate how linguistic features of clickbait vary across different news categories. These efforts aim to improve both the precision and recall of the model, leading to more robust and accurate clickbait detection in Chinese news headlines.

Acknowledgments

We are sincerely grateful to the three anonymous reviewers for their insightful and constructive feedback, which broadened our perspectives and enriched our understanding of the topic. Their thoughtful suggestions have been invaluable in refining this paper. While we were unable to incorporate all of their recommendations due to time constraints, their input will undoubtedly inform our future work and further research.

We also wish to express our deep appreciation to Professor I-Ping Wan and Professor Yu-Yun Chang for their guidance and support throughout the preparation of this manuscript. Their expertise and advice were instrumental in shaping the direction of this work. Any remaining limitations are solely our responsibility, and we look forward to building upon this foundation in subsequent studies.

References

Ángela Bazaco, Marta Redondo, and Pilar Sánchez-García. 2019. Clickbait as a strategy of viral journalism: conceptualisation and methods. *Revista Latina de Comunicación Social*, (74):94.

Iva Beleslin, Biljana Ratković Njegovan, and Maja S Vukadinović. 2017. Clickbait titles: Risky formula for attracting readers and advertisers. In *XVII International Scientific Conference on Industrial Systems*, volume 17, pages 364–369.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of pragmatics*, 76:87–100.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 9–16. IEEE.

Sarjak Chawda, Aditi Patil, Abhishek Singh, and Ashwini Save. 2019. A novel approach for clickbait detection. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)*, pages 1318–1321. IEEE.

Chi-Yen Chen and Wei-Yun Ma. 2017. Embedding wikipedia title based on its wikipedia text and categories. In *2017 International Conference on Asian Language Processing (IALP)*, pages 146–149. IEEE.

Chi-Yen Chen and Wei-Yun Ma. 2018. Word embedding evaluation datasets and wikipedia title embedding for chinese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Jhih-Sheng Fan, Mu Yang, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Hwe: Word embedding with heterogeneous features. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 39–46. IEEE.

Anna-Katharina Jung, Stefan Stieglitz, Tobias Kissmer, Milad Mirbabaie, and Tobias Kroll. 2022. Click me⋯! the influence of clickbait on user engagement in social media and the role of digital nudging. *Plos one*, 17(6):e0266743.

Tong Liu, Ke Yu, Lu Wang, Xuanyu Zhang, and Xiaofei Wu. 2021. Wcd: A new chinese online social media dataset for clickbait analysis and detection. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 368–372. IEEE.

Tong Liu, Ke Yu, Lu Wang, Xuanyu Zhang, Hao Zhou, and Xiaofei Wu. 2022. Clickbait detection on wechat: a deep model integrating semantic and syntactic information. *Knowledge-Based Systems*, 245:108605.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Xinyu Lun. 2021. The linguistic features of " clickbait " in chinese websites. In *2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021)*, pages 1976–1979. Atlantis Press.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 810–817. Springer.

Kate Scott. 2021. You won't believe what's in this paper! clickbait, relevance and the curiosity gap. *Journal of pragmatics*, 175:53–66.

Chaitanya Shinkhede. 2019. Digital frailty: Proliferation of clickbait, beguiled readers, and questioning the morality of online journalism. *marketing*, 6.

Bram Vijgen et al. 2014. The listicle: An exploring research on an interesting shareable new media phenomenon. *Studia Universitatis Babes-Bolyai-Ephemerides*, 59(1):103–122.