Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

A Study on an Integrated NER and Coreference Resolution Framework for Travel and Tourism Social Media Texts: Performance Comparison Using Advanced NLP Models

[1]Shraddha C. Kulkarni, [2]Ranjana S. Zinjore

[1]Assistant Professor, HPT ARTS and RYK science college, NASHIK, Maharashtra. India(91)

[2] Assistant Professor, MTES's Smt. G.G.Khadse College, Muktainagar, Dist. Jalgaon. Maharashtra. India(91)

Email: [1]dnyansam@gmail.com, [2]rszinjore14@gmail.com

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Named Entity Recognition, Coreference Resolution, Transformer-Based NLP, Social Media Text Analysis, Tourism Analytics.

Abstract

The study presents an integrated framework combining Named Entity Recognition (NER) and coreference resolution tailored for travel and tourism social media texts. Social media content is often noisy, informal, and multilingual, posing significant challenges for traditional NER methods. Leveraging advanced transformer based NLP models, including BERT and its variants, the framework addresses these challenges by improving entity extraction and contextual linking of references. This paper shows data collection, preprocessing techniques and model architectures adapted for the travel and tourism context. Evaluation on a manually curated social media dataset demonstrates notable improvements in entity recognition accuracy and coreference resolution consistency. The results highlight the potential of this approach for enhancing tourism analytics, trend monitoring, and recommendation systems. This work contributes to the growing body of research on applying deep learning techniques to domain specific social media text analysis in the tourism sector.

Introduction

The travel and tourism industry's growing reliance on digital content and user-generated data necessitates sophisticated language processing methods. NER is crucial for extracting structured data from unstructured sources to support recommendation engines, knowledge graphs, and personalized travel experiences. Early approaches ranged from lexicon-based to rule-based systems, but the field is now dominated by transformer-based and deep neural NLP models [1][2][3].

Large amount of data has been generated by social media platforms in unstructured form containing informal language, images, links, abbreviations, symbols, emojis, mix language text etc. Collecting such unstructured text and making it ready for further processing of research is a challenging task.

This paper proposes an integrated NLP framework for Named Entity Recognition (NER) combined with coreference resolution specifically designed for travel and tourism social media text. Unlike general domain datasets, tourism related social media content is highly noisy, informal, multilingual and context dependent. To address these challenges, the study leverages transformer based models, particularly BERT and its variants, along with a robust preprocessing pipeline that includes language detection, text normalization, tokenization, Pos tagging and domain adapted NER. The framework is evaluated on a manually curated tourism social media dataset, demonstrating improved entity extraction

accuracy and future implementation of better contextual linking of references. The work highlights the applicability of advanced NLP techniques for tourism analytics, trend detection and recommendation systems.

Research Objectives

1. To design a domain specific NLP framework for processing noisy and multilingual social media text in the travel and tourism sector.
2. To improve Named Entity Recognition performance for tourism related entities such as destinations, landmarks, and events using BERT based models.
3. To integrates coreference resolution with NER for better contextual understanding and entity coherence across social media posts.
4. To analyze the effectiveness of preprocessing techniques (language detection, text expansion, tokenization, POS tagging) for social media data.
5. To compare transformer based NER models with traditional approaches (CRF, BiLSTM) in the tourism domain.

Literature Review

The review synthesizes key papers and trends from the last two decades of research and details advances from machine learning to deep learning and transformer-based approaches, underscoring the current dominance and effectiveness of BERT and similar models for NER in travel and tourism. [1][2][4][5][6] Named Entity Recognition (NER) has long been recognized as an essential technique for extracting structured information such as destinations, facilities, and organizations from unstructured tourism content, including user reviews, travel blogs, and social media posts. [1][4] Early research in this area often relied on rule based and machine learning models, such as Conditional Random Fields (CRF), which performed well on structured or highly curated datasets but struggled with noisy, real world texts like social media. [1] CRF method is applied to tourism texts in Tamil and achieved an F1-score of 80.44%, highlighting the value of supervised learning in regional and multilingual tourism data. [7] With the advent of deep neural models, particularly the Bi-directional Long Short-Term Memory with CRF (BiLSTM CRF) combination, NER performance improved further, especially in modeling complex dependencies and context. [5] However, recent studies show that transformer-based architectures most notably BERT and its variants now set the benchmark for tourism NER, outperforming BiLSTM CRF and traditional machine learning models on both accuracy and generalization. [1] BERT model is used for extracting location, time, and names in Chinese tourism datasets, showing robust multilingual performance. [8] Most recent tourism NER studies use datasets scraped from travel focused websites and community forums, carefully annotated by experts to include categories such as heritage, natural, and purposefully built attractions. [1][6] The typical dataset sizes range from a few thousand to over ten thousand sentences, and entity labels differ slightly depending on region and application. [1][4] BERT-based NER models consistently yield F1-scores between 70% and 90%, depending on dataset complexity, annotation quality, and entity diversity. Performance increases when models are pretrained or domain-adapted to tourism-specific language and context. [1][2] Results from a recommendation engine using BERT-based NER demonstrated that automatic extraction of entities from web articles led to recommendations ranked highly by surveyed users (average interest level over 4 out of 5). [1] For travel and tourism named entity recognition using BERT and its variants, recent studies reports that BERT based models achieves strong performance with average F1-scores around 0.80 and precision values typically above 0.78 depending on the dataset and entity categories used.[1] Some specialized models and lightweight variants may trade some accuracy for efficiency. [2] A corpus for Dutch named Entity Recognition in archaeology domain is prepared and trained a Conditioned Random Field model on the dataset to assess the quality of NER with the data six new annotated entity types are identified which are not found in the general named entity recognition tasks. [9] A survey on deep learning-based solutions on Named Entity Recognition along with available NER resources including tagged NER corpora, off-the-shelf-NER systems with focus on NER in general domain and NER in English is discussed. [10] A model SEMFF-NER provides a method for Named Entity Recognition in social media texts that integrates multiscale features and syntactic information. The model also addresses the challenges, including entity sparsity and noise in social media texts. [11] Using data from Chinese newspapers and biographies, a new dataset that has been painstakingly annotated for named entities, entity linking, coreference, and relational information is presented. The collection presents significant study opportunities in the fields of digital humanities, linguistics, history, and natural language processing. In order to improve comprehension of the Chinese language and its historical context,

ongoing initiatives involve creating thorough annotations for event subtleties in newspaper data and creating NLP techniques especially suited to this particular historical data.[12] A neural network model using Transfer Learning (TL) for Part-of-speech (POS) tagging of social media texts is presented. Two scenarios of TL were experimented, the first is cross-domain TL, the second scenario is cross-task TL. [13] The of state-of-the-art POS taggers can significantly be improved for social media texts by taking training data from social media texts into account. A new social media text corpus WebTrain is created that contains 38,000 manually annotated tokens that can be used to retrain such taggers. Also, an adequate STTS annotation guideline for social media texts is proposed.[14] Text data preprocessing is one of the effective methods in terms of cleaning and making those unstructured data, structured and meaningful. Three different types of text data preprocessing technique (Stemming, Lemmatization and Spelling Correction) and its effect on sentiment produced is compared. An algorithm developed which can be used to provide proportional weight between the hashtag and cleaned text combined to obtain sentiment output. [15] The work presented to analyze the effectiveness and efficiency of different simple preprocessing tasks in transforming out-of-vocabulary tokens into in-vocabulary tokens over a database of Facebook comments in Spanish, gathered from 13 popular news portals. [16] A framework for processing and analyzing Arabic text on social media, with a focus on preprocessing and natural language processing (NLP) is presented. The proposed method successfully generates a structured dataset, thereby improving information extraction from Twitter data. It introduces novel approaches to cleaning, normalization, tokenization, stemming, and morphology generation. The method advances sentiment analysis and topic classification while highlighting the Arabic language's unique morphological aspects, which previous studies frequently overlooked. [17] TextPrep, a text preprocessing toolkit for topic modelling is presented, which demonstrate the value of good preprocessing in topic modelling. [18] Opinion mining techniques were used to classify a dataset of mobile application reviews that included slang, abbreviations, and jargons as positive, negative, or neutral. Pre-processing techniques were used to assess their impact on classifier performance, but contrary to expectations, they did not result in significant improvements. [19] A systematic reviews of preprocessing methods for sentiment analysis in Brazilian social media data is achieved. Research questions to outline the review's scope is established. Findings indicate a predominant focus on basic noise removal techniques, like stop word elimination and tokenization, although diverse preprocessing methods were noted. Nonetheless, the exploration of various approaches remains limited among authors.[20] A study on challenges of social media text processing and proposed solutions based on effective pre-processing is made. Also identified the sources of variations that lead to out-of-vocabulary (OOV) words and examined various pre-processing techniques, both traditional and application-specific. The implications of these techniques for text-processing applications were discussed.[21]

Methodology

This section details the methodology implemented to develop an NLP model for named entity recognition with coreference resolution. The model and architecture are proposed as illustrated in figure 1. The dataset is prepared by using manually curated social media text. The main aim of this work is to extract named entities and resolve its coreferences.

Proposed Architecture

The proposed architecture begins with data collection from social media platforms. This feeds into data preparation and cleaning, involving language detection, text normalization. The proposed architecture incorporates data augmentation as key phase including stemming, lemmatization and tokenization and POS tagging alongside prior techniques to enrich the tourism social media dataset for NER and coreference performance.

A pipeline implements:

1. Data Collection
2. Data Preparation
   a. Language Detection [1]
   b. Text Normalization [2]
   c. Tokenization and POS tagging using NLTK [7]
   d. Performance comparison of NER using fine-tuned BERT for tourism domain [1][5]
3. Implement a model for CorefResolve for coreference resolution of corresponding Named Entities.

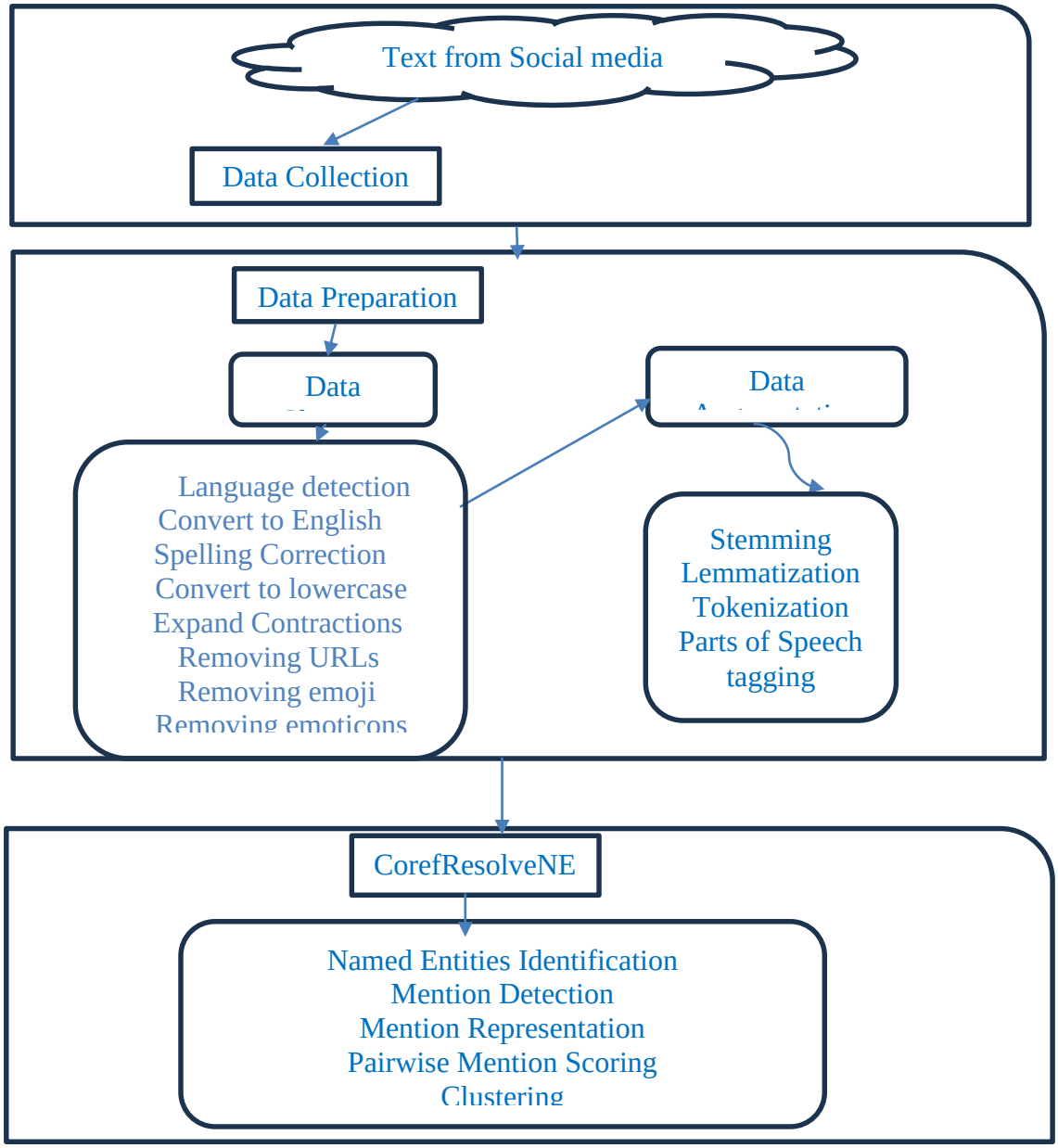

Figure 1. Proposed Architecture For Integrated Named Entity Recognition With Coreference Resolution

Data Collection

Data from social media related to travel is collected manually, then cleaned and manually annotated.[1] Datasets such as CONLL2003 and CONLL2012, though general purpose, provide useful baselines. A manual dataset is prepared containing social media texts with various attributes typical to the domain: mixed languages, abbreviations, and noisy text.

Table 1. A Sample Test Dataset prepared manually for implementation of above modules:

Text
I'll come back to you
Nashik'll be a vry nice place
Why MahaKumbh is only once in hundred years event?
When Sun and Moon together are in Capricorn and Jupiter is in Taurus (combination takes place once in twelve years) then Kumbh Mela is organised in Prayagraj
A blend of divine experiences and natural beauty! From the holy Trimbakeshwar Temple to
When twelve such Kumbh Melas called 'Purna Kumbh' are completed then comes Mahakumbh.

Algorithm 1: Language_detect (dataset1, languages)

Input: dataset1.xlsx" Contains a column named 'text' with textual data.
Output: "languages.xlsx" Same data as input, with an additional column 'Detected_Language' indicating the language of each text entry.

1) start
2) Load data
3) DATA-> Read(dataset1.xlsx)
4) Read data into variable text
5) call detect_language_safe(text):
6) If text is missing or not a string → return None.
7) Otherwise, attempt to detect the language.
8) If detection fails (e.g., error or empty result) → return None.
9) Return the detected language.
10) display result

Algorithm 2: Expand_text(dataset1, expanded_words)

Input: dataset1.xlsx" Contains a column named 'text' with textual data
Output: "expanded_text_output.xlsx" Same data as input, with an additional column expanded text indicating expanded text for contractions.

1) Start
2) Load data
3) Data -> Read(dataset1.xlsx)
4) Assign a sentence from data with contractions to variable text
5) call contractions.fix(text):
6) Join the expanded words into a single string expanded_text.
7) display expanded_text

Algorithm3: tokenization (dataset1, token)

Input: dataset1.xlsx" Contains a column named 'text' with textual data
Output: A new column 'tokens' in the same Excel file
1) Load data
2) DATA-> Read(dataset1.xlsx)
3) Assign a sentence to variable text
4) call Tokenization (text, token):
5) Apply Tokenization
6) Store the result in a new column 'token'

Algorithm3: POS_tagging (dataset1, POS_tag)

Input: dataset1.xlsx" Contains a column named 'text' with textual data
Output: A new column 'POS_Tags' in the same dataset
1) Load Excel file
2) Define POS tagging function
3) Tokenize the text into words
4) Call POS tagging(token,POS_tag):
5) Return the list of (word, tag) pair
6) Store the result in a new column 'POS_Tags'
7) Display result

The above modules are evaluated only for accuracy as a part of study. The proposed system architecture [figure 1] is designed for scalable processing of social media text in the travel and tourism domain, following these steps:

1. Data Collection: Table 1shows Social media posts focusing on content related to travel and tourism. Methods.
2. Data Preprocessing:
   a. Language Detection: Algorithm1 shows a robust language detection function is applied to identify the language for subsequent language specific processing.
   b. Text Expansion: Algorithm 2 shows an algorithm to expand English contractions to standard forms.
   c. Tokenization and POS Tagging: Algorithm 3 and 4 shows algorithm for tokens extraction and tagging using libraries like NLTK.
   d. Named Entity Recognition: Pre-trained NER models, possibly fine-tuned on tourism datasets, extract location names, landmarks, and organization mentions.
   e. Coreference Resolution: This step facilitates understanding of entity mentions across a text, improving recognition coherence.

Several Python based algorithms were developed for these tasks, operating on an Excel-based dataset extracted from social media platforms containing approximately 100 example records for testing.

Table 5. Performance Evaluation

Module Name	Accuracy
Language Detection Module	0.6408%
Expansion Text Module	0.9717%

For current study of "Developing a framework for Named entity recognition with coreference resolution in social media text content" BERT model and its variants are studied as follows.

BERT Models

BERT model, a significant milestone in the development of NLP is a model that uses stack of transformer encoders and is trained using a novel approach called masked LM. Masked LM involves randomly selecting 15% of input tokens and then: i) Hiding 80% of them. ii) Replacing 10% with random token .iii) Leaving the remaining 10% unchanged. The model's goal is to predict the original modified tokens, which forces it to maintain a distributional contextual representation of each input token.[3]

RoBERTa utilizes a language masking strategy to predict hidden sections of text during pretraining. It modifies several key hyperparameters of BERT, notably by removing the next sentence pretraining objective and employing significantly larger mini batches and learning rates. These changes enhance the masked language modelling objective compared to BERT leading to improved performance on various downstream NLP tasks. Tuning the BERT training procedure A DONE IN RoBERTa, significantly boosts performance and demonstrates the potential of self supervised training techniques to rival or exceed traditional supervised methods.

Tourism NER

Social media serves as a massive repository of information rich with personal stories, comments, and shared knowledge. This content enables travelers who are unsure of where to visit to quickly find recommendations in the tourism sector. A simple query, such as typing "places to visit in Bali" can bring up many blog articles to assist in decision making. However, reading all this information without a helper can be overwhelming. To address this challenge, a Bidirectional Encoder Representation from transformer based tourism named entity recognition system was developed. This system is used to highlight specific tourist destination places in the query result. BERT is a state-of-the-art machine learning framework for natural language processing. The existing Tourism NER model specifies three types of tourist destination's heritage, Natural and purposefully built (manmade or artificial) within social media posts and articles. The model achieved a respectable average F1-score of 0.80 and has been integrated into a functional recommendation system.

Comparison of Named Entity Recognition Using Specialized NLP Models

Named Entity Recognition (NER) models applied to travel and tourism have evolved from traditional machine learning approaches to sophisticated deep learning and transformer-based architectures. The most prominent specialized NLP models in this domain include Conditional Random Fields (CRF), BiLSTM-CRF, and various BERT-based models with domain adaptation

Table 6. Comparison of Named Entity Recognition Using Specialized NLP Models

Model	Precision	F1-Score
BERT based cased	0.81[1]	0.81[1][2]
BERT based uncased	0.79[1]	0.78[1]
TourBERT	0.83[4]	0.82[4]
BiLSTM-CRF	0.72[1]	0.75[1]
DistilBERT	~0.78[3]	~0.77[3][5]
DCM-BERT	0.84[6]	0.85[6]

This comparison highlights the advantages of specialized NLP transformer models like BERT for NER in travel and tourism and provides a clear rationale for their adoption in research projects and practical systems involving social media text. [FIGURE 2]

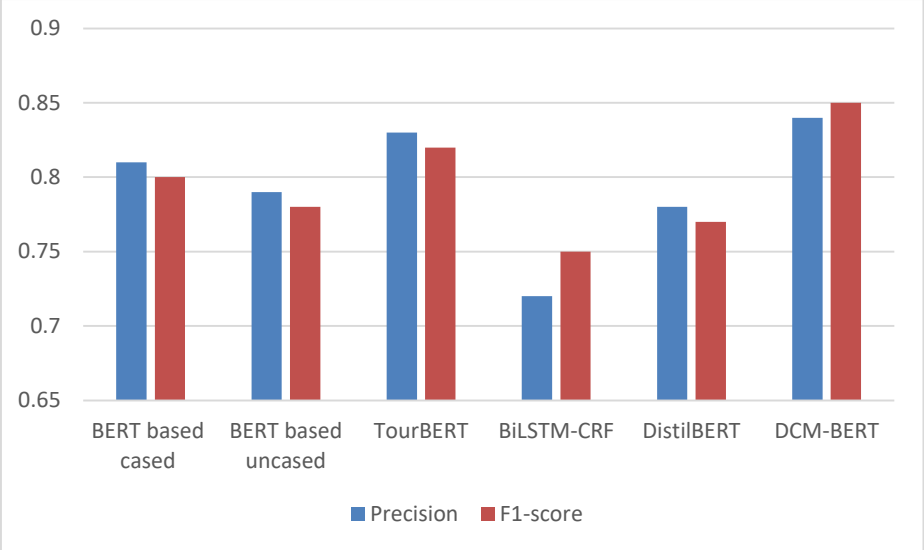

Figure 2. Performance Evaluation of BERT Variants

Observations

BERT and its variants have become the de facto standards in travel and tourism NER due to their superior handling of contextual nuances, multilingual data, and noisy social media text. They outperform traditional CRF and BiLSTM CRF by a significant margin (often 5-15% or more in F1-score) in tourism entity recognition tasks.

Challenges

Social media's informal language code switching, slang, abbreviations, and emojis complicates cleaning and normalization for NER and coreference resolution. Largescale, real time data collection requires effective filtering like geotagging and keywords to ensure relevance. Limited annotated datasets for tourism-specific NER and coreference resolution demand extensive manual labelling.

Results And Discussions

To develop an integrated named entity recognition and coreference resolution framework, a specialized architecture was proposed. BERT models and their variants were studied to assess their performance for this task. Performance metrics for these models were reviewed and compared. According to the architecture, text normalization algorithms were implemented and evaluated for accuracy; the language detection module achieved an accuracy of 64.08%, while the text expansion module reached 97.17%. The coreference resolution module remains to be implemented, as it presents substantial challenges due to the ambiguity and unstructured nature of social media text, particularly when resolving different named entities and their references.

Conclusion

State-of-the-art NLP models like BERT are highly effective for tourism NER, outperforming previous techniques and supporting a wide range of downstream tourism technology applications. Future work will focus on addressing multilingual data, low-resource settings, and expanding entity types relevant to emerging travel trends [1][2][5].

This research presents a novel framework for NER in the travel and tourism domain, focusing on the nuanced challenges posed by social media text. Through effective preprocessing and leveraging specialized NLP models, the system promises enhanced extraction of tourism-related entities, supporting applications in tourism analytics, real-time event monitoring, and personalized recommendations.

Future Work

Remaining work includes refining the architecture, extending data collection to larger datasets, developing and implementing specialized NER and coreference resolution algorithms using hybrid models, and comprehensive evaluation of system performance on real world social media posts.

Acknowledgement

The author expresses her gratitude to her research supervisor Dr. Ranjana S. Zinjore Madam, Dr. Manoj Patil Sir, Dr. Varsha. M. Pathak Madam for their valuable suggestions throughout the present study.

References

[1] Fudholi, D. H., Zahra, A., Rani, S., Huda, S. N., Paputungan, I. V., & Zukhri, Z. (2023). BERT-

based tourism named entity recognition: making use of social media for travel recommendations. PeerJ Computer Science, 9. https://doi.org/10.7717/PEERJ-CS.1731

[2] Bouabdallaoui, I., Guerouate, F., Bouhaddour, S., Saadi, C., & Sbihi, M. (2021). Named Entity Recognition applied on Moroccan tourism corpus. Procedia Computer Science, 198, 373–378. https://doi.org/10.1016/j.procs.2021.12.256

[3] Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. International Journal of Geographical Information Science, 37(4), 747–766. https://doi.org/10.1080/13658816.2022.2133125

[4] Arefieva, V., & Egger, R. (n.d.). TourBERT: A pretrained language model for the tourism industry. https://www.kaggle.com/jiashenliu/515k-

[5] Abadeer, M. (2020). Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts. https://github.com/huggingface/

[6] Gajula, S. (2024). Cybersecurity risk prediction using graph neural networks. Journal of Information Systems Engineering and Management, 9(4S), 3301–3315.

[7] Vijayakrishna R, & Sobha L. (2008). Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields.

[8] Hu, Y., Nuo, M., & Tang, C. (2019). A Deep Learning Approach for Chinese Tourism Field Attribute Extraction. Proceedings - 2019 15th International Conference on Computational Intelligence and Security, CIS 2019, 108–112. https://doi.org/10.1109/CIS.2019.00031

[9] Brandsen, A., Verberne, S., Wansleeben, M., & Lambers, K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. 11–16. https://doi.org/10.5281/zenodo.3544544

[10] Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

[11] Li, Y., Zhou, Y., Hu, X., Li, Q., & Tian, J. (2024). A method for named entity recognition in social media texts with syntactically enhanced multiscale feature fusion. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-78948-5

[12] Blouin, B., Armand, C., & Henriot, C. (n.d.). A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers. https://gitlab.com/enpchina/ENP-NER

[13] Meftah, S., Semmar, N., & Sadat, F. (2018). A neural network model for part-of-speech tagging of social media texts A neural network model for part-of-speech tagging of social media texts A Neural Network Model for Part-Of-Speech Tagging of Social Media Texts. In ELRA. https://cea.hal.science/cea-04572171v1

[14] Reyer, M., Mathar, R., Neunerdt, M., & Trevisan, B. (n.d.). Part-Of-Speech Tagging for Social Media Texts. https://doi.org/10.13140/2.1.2905.7284

[15] Mothe, Josiane., Son, L. Hoang., & Nguyen, T. Q. Vinh. (2019). Proceedings of 2019 11th International Conference On Knowledge And Systems Engineering : KSE 2019 : October 24-26, 2019, Da Nang, Vietnam. IEEE.

[16] Tessore, J. P., Esnaola, L. M., Russo, C. C., & Baldassarri, S. (2019, June 25). Comparative analysis of preprocessing tasks over social media texts in Spanish. ACM International Conference Proceeding Series. https://doi.org/10.1145/3335595.3335632

[17] Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic text on social media. Heliyon, 7(2). https://doi.org/10.1016/j.heliyon.2021.e06191

[18] Churchill, R., & Singh, L. (2021). textPrep: A text preprocessing toolkit for topic modeling on social media data. Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021, 60–70.

https://doi.org/10.5220/0010559000600070

[19] dos Santos, F. L., & Ladeira, M. (2014). The role of text pre-processing in opinion mining on a social media language dataset. Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014, 50–54. https://doi.org/10.1109/BRACIS.2014.20

[20] Cirqueira, D., Fontes Pinheiro, M., Jacob, A., Lobato, F., & Santana, A. (2019). A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media. Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018, 746–749. https://doi.org/10.1109/WI.2018.00008

[21] Khan, J., Ahmad, K., Jagatheesaperumal, S. K., & Sohn, K. A. (2025). Textual variations in social media text processing applications: challenges, solutions, and trends. Artificial Intelligence Review, 58(3). https://doi.org/10.1007/s10462-024-11071-z