Term Extraction using Machine Learning Techniques

Temirlan Zhexembayev
BDA-2205

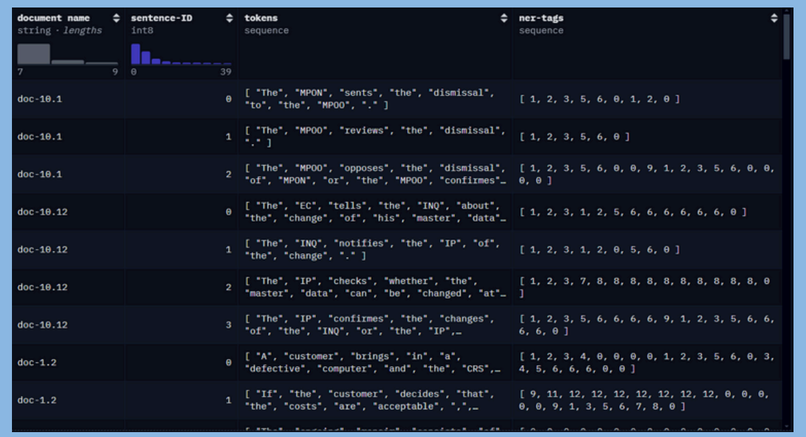
Abstract

Aspect term extraction (ATE) is crucial for sentiment analysis and text processing. This study uses the PET dataset and applies machine learning and deep learning techniques for extracting key terms. The findings highlight the effectiveness of transformer models compared to traditional approaches.

Introduction

01.

•Aspect term extraction (ATE) identifies key concepts in text data. This project explores different methodologies, including statistical, rule-based, and deep learning models to improve ATE accuracy using the PET dataset.



document name string · *lengths*	sentence-ID int8	tokens sequence	ner-tags sequence
	7 9	0 39	
doc-10.1	0	["The", "MPON", "sents", "the", "dismissal", "to", "the", "MPOO", "."]	[1, 2, 3, 5, 6, 0, 1, 2, 0]
doc-10.1	1	["The", "MPOO", "reviews", "the", "dismissal", "."]	[1, 2, 3, 5, 6, 0]
doc-10.1	2	["The", "MPOO", "opposes", "the", "dismissal", "of", "MPON", "or", "the", "MPOO", "confirmes"…	[1, 2, 3, 5, 6, 0, 0, 9, 1, 2, 3, 5, 6, 0, 0, 0, 0]
doc-10.12	0	["The", "EC", "tells", "the", "INQ", "about", "the", "change", "of", "his", "master", "data"…	[1, 2, 3, 1, 2, 5, 6, 6, 6, 6, 6, 6, 0]
doc-10.12	1	["The", "INQ", "notifies", "the", "IP", "of", "the", "change", "."]	[1, 2, 3, 1, 2, 0, 5, 6, 0]
doc-10.12	2	["The", "IP", "checks", "whether", "the", "master", "data", "can", "be", "changed", "at"…	[1, 2, 3, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 0]
doc-10.12	3	["The", "IP", "confirmes", "the", "changes", "of", "the", "INQ", "or", "the", "IP",…	[1, 2, 3, 5, 6, 6, 6, 6, 9, 1, 2, 3, 5, 6, 6, 6, 0]
doc-1.2	0	["A", "customer", "brings", "in", "a", "defective", "computer", "and", "the", "CRS",…	[1, 2, 3, 4, 0, 0, 0, 0, 1, 2, 3, 5, 6, 0, 3, 4, 5, 6, 6, 6, 0, 0]
doc-1.2	1	["If", "the", "customer", "decides", "that", "the", "costs", "are", "acceptable", ",",…	[9, 11, 12, 12, 12, 12, 12, 12, 12, 0, 0, 0, 0, 0, 9, 1, 3, 5, 6, 7, 8, 0]

Literature Review

Article	Methodology	Strengths	Weaknesses	Knowledge Gap
Aspect term extraction and optimized deep learning for sentiment classification	Squirrel Search Mayfly Algorithm (SSMA) combined with Hierarchical Deep Learning (HDLTex) on Amazon reviews.	High precision (0.936), F-measure (0.937), and recall (0.941). Innovative optimization with SSMA.	Limited to Amazon reviews; applicability to other domains not discussed.	Need for real-world testing across diverse datasets.
Homonym and polysemy approaches in Indonesian-English translation	Morphology extraction with BERT embeddings, NER, and semantic similarity for handling homonymy and polysemy.	Innovative handling of linguistic features like homonymy and polysemy. Improves translation accuracy.	No quantitative comparison with existing translation systems. Relies heavily on domain-specific linguistic rules.	Further exploration of multilingual and low-resource language scenarios.

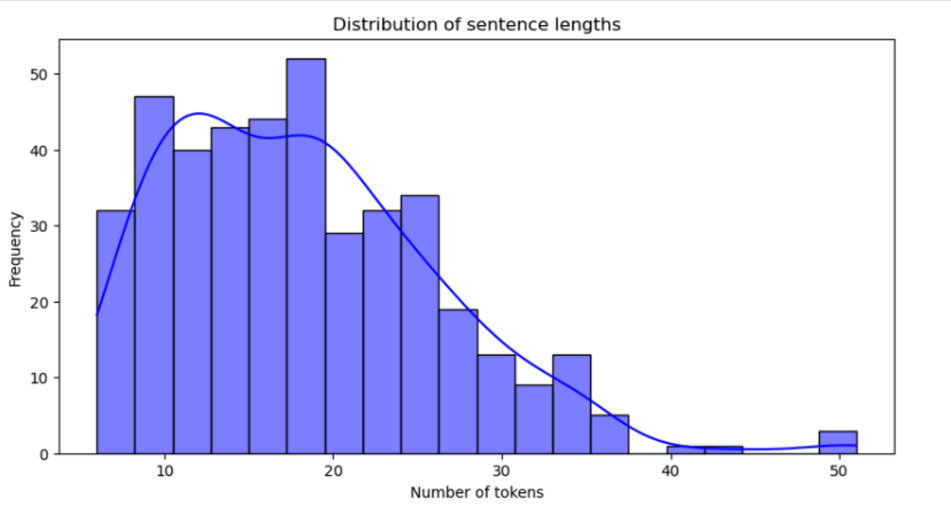
Literature Review

Article	Methodology	Strengths	Weaknesses	Knowledge Gap
Cross-domain aspect term extraction with character-level features	Multi-channel encoder incorporating character-level, local, and global features to reduce dependency on labeled data.	Reduces reliance on labeled data. Achieves state-of-the-art results on benchmark datasets.	Effectiveness in real-world cross-domain applications remains unclear. Limited evaluation across vastly differing domains.	Extending beyond syntactic similarities for knowledge transfer.
Sentiment analysis in aspect term extraction for mobile phone tweets	NLTK for multi-aspect term extraction combined with machine learning (Random Forest, KNN, SVM) for sentiment classification of iPhone and Samsung tweets.	Use of real-world Twitter data. Combines deep learning and traditional ML methods.	Focuses on specific brands (iPhone, Samsung); may lack generalizability.	Generalization to other social media platforms or industries.

Literature Review

Article	Methodology	Strengths	Weaknesses	Knowledge Gap
Tone or term: Machine-learning text analysis in bond pricing	Machine-learning techniques to extract featured vocabulary and text analysis scores for credit rating reports in China's bond markets.	Enhanced vocabulary coverage, reduces misclassification, and mitigates equal-weighting issues in BoW methods.	Dataset limited to China's bond markets. May lack applicability outside the financial domain.	Broader exploration of financial text analytics in other global markets.
Aspect term extraction from multi-source domain using E-LDA	Enhanced Latent Dirichlet Allocation (E-LDA) model for topic modeling and aspect extraction across multi-source domains.	High coherence score (0.5727). Captures domain-specific sentiments and aspects effectively.	Focus on coherence score; lacks sentiment-specific performance metrics like precision or recall.	Further exploration of sentiment classification across multi-source domain datasets.

Methodology



- Dataset: PETv11 (Token Classification)
- Data Preprocessing: Tokenization, Stopword Removal, Lemmatization, Named Entity Recognition (NER)
- Models: SVM, Random Forest, LDA, BERT, RoBERTa
- Evaluation: Precision, Recall, F1-Score, Coherence Score

```
🔹 Top TF-IDF Words: ['service', 'report', 'process', 'mpon', 'customer']

🔹 LDA Topics:
(0, '0.125*"the" + 0.052*"." + 0.033*"of" + 0.028*"The" + 0.022*","')
(1, '0.086*"the" + 0.045*"." + 0.035*"," + 0.026*"to" + 0.023*"is"')

🔹 BERT Aspect Classification: ['AI', 'Machine Learning', 'AI', 'Data Science', 'AI',
I', 'Machine Learning', 'AI', 'AI', 'AI', 'AI', 'AI', 'AI', 'AI', 'AI', 'Machine Learni
```

Results & Discussion

	precision	recall	f1-score	support
1	0.51	0.25	0.34	83
5	0.41	0.47	0.44	87
6	0.49	0.68	0.57	214
8	0.00	0.00	0.00	71
11	0.00	0.00	0.00	13
3	0.76	0.58	0.66	96
2	0.56	0.56	0.56	113
12	0.41	0.09	0.15	78
0	0.66	0.82	0.73	687
4	0.00	0.00	0.00	15
9	0.00	0.00	0.00	22
7	0.00	0.00	0.00	15
10	0.00	0.00	0.00	1
13	0.00	0.00	0.00	1
14	0.00	0.00	0.00	0
micro avg	0.60	0.60	0.60	1496
macro avg	0.25	0.23	0.23	1496
avg	0.54	0.60	0.55	1496

- TF-IDF: Efficient but lacks context awareness
- NER (SpaCy): Accurate but dataset-dependent
- LDA: Provides themes but needs manual interpretation
- Transformers: Best performance with contextual understanding

Conclusion & Future Work

This study demonstrates the importance of combining traditional and deep learning methods for ATE.

Future work includes domain-specific fine-tuning, incorporating syntactic analysis, and using multimodal data integration for improved term extraction.

By leveraging these advanced techniques, researchers can enhance the accuracy and efficiency of automatic term extraction processes. Additionally, exploring collaborative frameworks that involve interdisciplinary expertise may further refine the models and broaden their applicability across diverse fields. As technology evolves, the integration of real-time feedback mechanisms and adaptive learning algorithms could also play a crucial role in continuously optimizing the performance of ATE systems. Ultimately, the goal is to create robust, scalable solutions that can seamlessly handle the ever-growing complexity of terminological data in various domains.

References

[1] Adilakshmi, K. Aspect term extraction and optimized deep learning for sentiment classification. (2024) Social Network Analysis and Mining, 14 (1), art. no. 221. https://www.scopus.com/record/display.uri?eid=2-s2.0-85210104156&origin=resultslist&sort=plf-f&src=s&sot=b&sdt=b&cluster=scosubtype%2C%22ar%22%2Ct&s=TITLE%28term+AND+extraction%29&sessionSearchId=9bd5a89d2be6ee84f549410b2d16b470&relpos=10

[2] Dhanal R.J., Ghorpade,V. Aspect term extraction from multi-source domain using enhanced latent Dirichlet allocation. (2024) Indonesian Journal of Electrical Engineering and Computer Science, 35 (1), pp. 475-484. https://www.scopus.com/record/display.uri?eid=2-s2.0-85192060384&origin=resultslist&sort=plf-f&src=s&sot=b&sdt=b&cluster=scosubtype%2C%22ar%22%2Ct&s=TITLE%28term+AND+extraction%29&sessionSearchId=9bd5a89d2be6ee84f549410b2d16b470&relpos=45

[3] Harjo, B., Muljono, Abdullah, R. Homonym and polysemy approaches with morphology extraction in weighting terms for Indonesian to English machine translation. (2024) International Journal of Electrical and Computer Engineering, 14 (6), pp. 7036-7045. https://www.scopus.com/record/display.uri?eid=2-s2.0-85206088645&origin=resultslist&sort=plf-f&src=s&sot=b&sdt=b&cluster=scosubtype%2C%22ar%22%2Ct&s=TITLE%28term+AND+extraction%29&sessionSearchId=9bd5a89d2be6ee84f549410b2d16b470&relpos=12

[4] Naramula, V., A, K. Sentiment analysis in aspect term extraction for mobile phone tweets using machine learning techniques (2024) International Journal of Pervasive Computing and Communications, 20 (4), pp. 1-20. http://www.emeraldinsight.com/products/journals/journals.htm?id=ijpcc

[5] Peng, Y., Shi, L., Shi, X., Tan, S. Tone or term: Machine-learning text analysis, featured vocabulary extraction, and evidence from bond pricing in China (2024) Journal of Empirical Finance, 78, art. no. 101534. https://www.sciencedirect.com/science/journal/09275398

[6] Wang, D., Li, W. Cross-domain aspect term extraction incorporating character-level features. (2024) Computers and Electrical Engineering, 120, art. no. 109653. https://www.scopus.com/record/display.uri?eid=2-s2.0-85203147076&origin=resultslist&sort=plf-f&src=s&sot=b&sdt=b&cluster=scosubtype%2C%22ar%22%2Ct&s=TITLE%28term+AND+extraction%29&sessionSearchId=9bd5a89d2be6ee84f549410b2d16b470&relpos=15

Thank you for your attention