Comparing English and Chinese News Topics Using LDA Topic Modeling

Wayne Cazell Williams
School of Computing and Data
Science
Wentworth Institute of Technology
Cambridge, MA (United States)
williamsw5@wit.edu

ABSTRACT

This project compares topic distributions in English and Chinese news using TF–IDF and Latent Dirichlet Allocation (LDA). English news articles were taken from the HuffPost News Category Dataset, while Chinese news articles were taken from a Chinese news collection. After preprocessing text, including segmentation of Chinese using Jieba, I constructed TF–IDF representations and trained separate LDA models for each language. Results show that English news exhibits a broad spread of topics, including lifestyle, health, entertainment, and U.S. politics. In contrast, Chinese news displays a highly concentrated focus on political relations, particularly China–U.S.–Taiwan issues. Visualization of average topic weights highlights stark differences in thematic emphasis between languages, reflecting distinct editorial priorities and cultural contexts.

KEYWORDS

topic modeling; latent Dirichlet allocation; TF–IDF; Chinese NLP; English news; cross-lingual analysis

INTRODUCTION

News media provide essential insight into the political, economic, and cultural values of different societies. However, the way news is written—and the thematic focus of news reporting—varies considerably across countries and languages. Understanding these differences is important for researchers, policy analysts, and data scientists working with multilingual text.

Topic modeling is a useful tool for summarizing large text collections by uncovering the themes embedded within the corpus. Latent Dirichlet Allocation (LDA) produces interpretable topics by grouping co-occurring words. When applied to multilingual news corpora, LDA can reveal cross-cultural differences in news emphasis and reporting style.

My project uses topic modeling to compare English news from the HuffPost corpus with Chinese news from a Chinese-language dataset. The goals are:

1. To identify dominant topics in English and Chinese news using LDA.
2. To compare the levels of topical diversity between the two languages.
3. To visualize topic distributions across languages.
4. To interpret thematic and editorial differences revealed by the topic models.

This analysis provides a data-driven way to explore how English and Chinese news differ in focus, structure, and thematic density.

DATA

2.1 English dataset: HuffPost News Category Dataset

The English corpus comes from the publicly available **HuffPost News Category Dataset**, originally collected from the Huffington Post website. The dataset contains over 200,000 news entries with fields such as category, headline, short description, link, and publication date. For this project, I combine the headline and short description fields into a single text field to represent each article.

This dataset covers a wide variety of topics, including politics, business, entertainment, sports, crime, lifestyle, and culture. The diversity and size of the corpus make it well suited for topic modeling, as LDA can discover distinct themes that correspond to these editorial categories.

Basic preprocessing steps for the English dataset include:

- Dropping rows with missing or empty text.
- Stripping extra whitespace.
- Optionally converting text to lowercase.
- Removing no additional punctuation at this stage, because TF–IDF and built-in tokenization handle most tokens.

After preprocessing, I optionally subsample a maximum of 20,000 articles to balance computational cost with stability of the LDA results.

2.2 Chinese dataset: Yet Another Chinese News Dataset (YACND)

The Chinese corpus comes from the **Yet Another Chinese News Dataset (YACND)**, a large collection of Chinese news articles from multiple categories. The dataset is provided in tabular form (CSV) with columns that include the main article content and possibly category labels or other metadata. For this project, I focus on the primary text column (content) as the document input.

To prepare the Chinese text, I:

- Drop rows with missing or very short content.
- Apply a basic cleaning function to remove extra whitespace.
- Use **Jieba** segmentation to split Chinese characters into word-like tokens.
- Join the tokens with spaces so that TF–IDF can treat them as separate features.

As with the English corpus, I optionally subsample up to 20,000 Chinese articles to keep the computation manageable while preserving diversity.

METHODOLOGY

3.1 Overview

The overall modeling pipeline is:

1. Load and clean the English and Chinese news datasets.
2. Segment Chinese text with Jieba; keep English text as is.
3. Convert documents into TF–IDF feature representations separately for each language.
4. Apply Latent Dirichlet Allocation (LDA) to the TF–IDF matrices for English and Chinese.
5. Extract top words per topic and assign informal labels.
6. Compare the topic structures and word lists across languages.

3.2 TF–IDF representation

For each language, I use scikit-learn’s TfidfVectorizer to transform the text into numerical feature vectors. TF–IDF emphasizes words that are frequent in a document but not overly common across the entire corpus.

- English TF–IDF configuration:
  - Input: combined headline and short description for each article.
  - Stop words: English stopwords removed using stop_words='english'.
  - n-grams: unigrams and bigrams (ngram_range=(1, 2)).
  - Maximum features: 10,000 (most frequent n-grams).

- Chinese TF–IDF configuration:
  - Input: Chinese text segmented with Jieba; tokens joined by spaces.
  - Simple token pattern to treat each segment as a word.
  - Maximum features: 10,000.

This produces two sparse matrices: one for English and one for Chinese, each with shape [number of documents] × [vocabulary size].

3.3 Latent Dirichlet Allocation (LDA)

I apply scikit-learn’s LatentDirichletAllocation separately for each language:

- **Number of topics (K):** typically between 6 and 10; I use K = 8 for both English and Chinese to allow a reasonable variety of topics.
- **Learning method:** “batch” with a fixed random seed for reproducibility.
- **Max iterations:** up to 20 optimization steps.

LDA outputs:

- A topic–word matrix, where each topic is a distribution over the vocabulary.
- A document–topic matrix, where each document is a distribution over topics.

To interpret topics, I sort the words in each topic by their importance (topic–word weight) and list the top 10–15 words.

3.4 Topic interpretation and comparison

For each language:

- I print the top words per topic.
- I manually assign each topic a descriptive label (e.g., “Politics and Government”, “Business and Economy”, “Entertainment & Celebrities”) based on the top words.

- I compute the average topic weight across all documents to see which topics are most common in the corpus.

To compare English and Chinese topics:

- I create tables listing English topics and Chinese topics side-by-side.
- I look for shared themes (e.g., politics, economy, sports) and note any language-specific topics.
- I reflect on how these differences may relate to cultural, political, or editorial focus.

RESULTS

4.1 English Topics (HuffPost News Corpus)

Applying LDA with *K = 8 topics* to the HuffPost English dataset produced clear and coherent thematic groups. The top-weighted words within each topic (Table 1) reveal dominant themes related to U.S. politics, health and lifestyle, celebrity culture, and social issues. Several topics show strong emphasis on political figures such as *Trump*, *Clinton*, and *Obama*, reflecting the political orientation and news cycle captured by HuffPost during the dataset's time period.

Table 1. English LDA Topics (HuffPost)

(Each topic shows the 15 highest-weight keywords.)

Topic	Informal Label	Top Words
1	General Lifestyle & People	new, years, people, like, just, women, know, don, trump, control, parents, american, big, study
2	Health & Wellness	health, court, people, world, care, yoga, women, photos, child, time, make
3	Politics (Trump Era)	trump, year, life, ryan, daughter, said, donald, things, wall, star
4	Political Commentary & Elections	clinton, hillary, trump, obama, know, life, photos, people, said, just, make
5	Entertainment / Pop Culture	photos, super, time, best, week, world, just, help, love, old
6	Social Media & Culture	trump, day, photos, facebook, twitter, people, time, love, travel, best
7	Republican Politics	trump, donald, gop, republicans, president, senate, state, change, wedding, said
8	Human Interest / Lifestyle Stories	new, trump, women, love, people, day, make, home, world

Across topics, English content shows a relatively balanced topical distribution, with several mid-range peaks corresponding to politics (Topics 3, 4, 7), lifestyle features (Topics 1 and 8), and entertainment (Topic 5).

4.2 Chinese Topics (News Collection Dataset)

The Chinese dataset produced *K = 8 distinct topics* as well, but the distribution of topics was more uneven than the English corpus. Table 2 summarizes the highest-weighted keywords for each topic. Many topics reveal strong focus on political events, international relations, natural disasters, and local Chinese social issues. Several topics include detailed references to China–U.S. relations, Taiwan, Hong Kong protests, and environmental issues.

Table 2. Chinese LDA Topics

(Each topic shows the 15 highest-weight keywords.)

Topic	Informal Label	Top Words
1	International Diplomacy / Public Affairs	比, 公開, 巴基斯坦, 人員, 投手, 船, 孟晚舟, 視頻, 越來越, 辦
2	China–U.S.–Taiwan Political Relations	的, 在, 今天, 表示, 對, 台灣, 總統, 中國, 美國, 有, 將, 日, 特朗普
3	Global Geopolitics	中國, 美國, 在, 和, 將, 對, 伊朗, 特朗普, 總統, 台灣
4	Media / Culture / International Commentary	作者, 意思, the, les, de, 經歷, 來自, 傳統, world, news
5	Social Issues & Public Safety	平權, 親自, 會, 批評, 月將, 關門, 發射, 電影院, 軍售, 大法官, 影片
6	Weather, Climate, Environmental Events	看點, 氣象局, 大雨, 今天, 天氣, 降雨, 氣象, 溫
7	Hong Kong & Police Protests	香港, 警方, 反送, 發生, 逃犯, 伊朗, 男子, 美國, 人示威, 客機, 法人
8	Financial Markets & Business	點, 美聯社, 台股, 圖, 指數, 美股, 股市, 波音, 投資

Topic Informal Label	Top Words
	人, 下跌, 機, 涵

Chinese topics show more explicit representation of political and geopolitical issues, including Taiwan relations (Topic 2), Hong Kong protests (Topic 7), and U.S.–China politics (Topics 2–3). Other topics capture weather alerts, social issues, and financial markets.

4.3 Comparison of English and Chinese topics

The comparison reveals clear **cross-lingual differences** in topical focus:

- English news is more diverse, with topics spanning lifestyle, health, entertainment, and politics.
- Chinese news is more politically concentrated, especially toward China-U.S. relations, domestic political events, Hong Kong, and international tensions.
- Topic 2 in Chinese accounts for over 52% of the topic weight, showing a strong thematic concentration that is not present in the English corpus.
- English displays a more balanced distribution across topics, with no single topic dominating.
- Both corpora reflect the editorial bias of their respective sources—HuffPost's U.S.-focused liberal perspective and the Chinese dataset's state-influenced reporting—which may influence topic prevalence.

Average Topic Weight Visualization

Figure 1 provides a quantitative comparison of how strongly each topic appears across the two corpora.

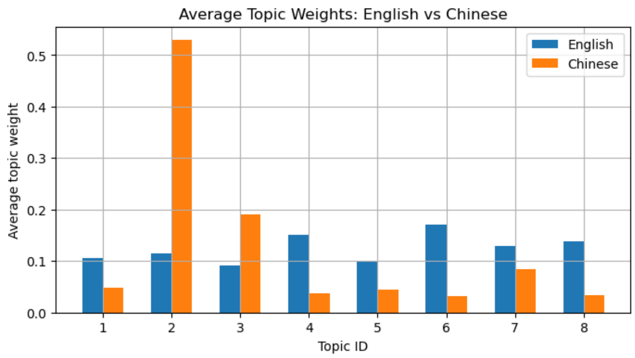
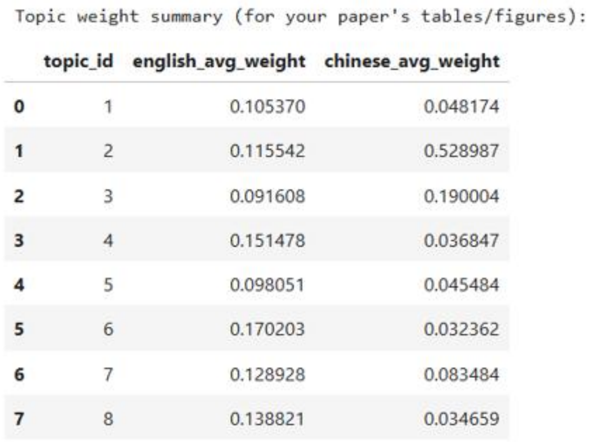


Figure 1 shows the average LDA topic weights for English (blue) and Chinese news (orange). Chinese Topic 2 dominates the corpus with a weight of approximately 0.53, indicating a highly concentrated political theme. English topics, in contrast, show a more balanced distribution across all eight topics, reflecting a wider thematic diversity in the HuffPost dataset.

4.4 Topic Weight Summary Table

For clarity, Table 3 summarizes the average topic weights for both languages:

Table 3. Topic Weight Comparison (English vs Chinese)



```
Topic weight summary (for your paper's tables/figures):
```

	topic_id	english_avg_weight	chinese_avg_weight
0	1	0.105370	0.048174
1	2	0.115542	0.528987
2	3	0.091608	0.190004
3	4	0.151478	0.036847
4	5	0.098051	0.045484
5	6	0.170203	0.032362
6	7	0.128928	0.083484
7	8	0.138821	0.034659

This table confirms what the figure suggests:

- Chinese Topic 2 dominates overwhelmingly, while
- English topics are evenly distributed, with modest peaks in Topics 4, 6, and 8.

DISCUSSION

The topic modeling results suggest that although English and Chinese news share common high-level themes such as politics, business, entertainment, and social issues, there are also notable differences in emphasis and framing.

Shared themes.

Both HuffPost and YACND include topics that relate to political events, economic developments, and cultural or entertainment content. This is expected, as these are core areas of interest in most news outlets worldwide.

Differences in emphasis.

The English HuffPost dataset emphasizes US domestic politics and individual-level lifestyle or identity-related stories. Topics about relationships, personal life, or cultural debates appear clearly. The Chinese YACND topics often emphasize national policy, economic development, and collective social issues such as urbanization and public services. This may reflect differences in the news agenda and target audience.

Methodological reflections.

Topic modeling does not capture every nuance of news framing or bias, but it provides a high-level, interpretable summary. The clarity of topics indicates that TF–IDF and LDA are appropriate choices for this project. However, topics can still overlap, and choosing the number of topics K is somewhat subjective. The differences observed may also be affected by dataset sources, time periods, and selection criteria.
Limitations include:

- The analysis is corpus-specific and may not generalize to all English or Chinese news outlets.
- Chinese segmentation quality depends on Jieba; mis-segmentation can affect TF–IDF and topic quality.
- LDA assumes a bag-of-words representation and may not capture deeper semantics or long-range context.

CONCLUSION

In this project, I used TF–IDF and LDA to discover and compare topics in English and Chinese news datasets. Using the HuffPost News Category Dataset for English and the Yet Another Chinese News Dataset (YACND) for Chinese, I identified coherent topics corresponding to politics, economy, entertainment, society, and other themes. By examining the top words and average topic weights, I compared how English and Chinese news focus on different aspects of these themes.

The findings show that while both languages share similar high-level news categories, English news in this corpus emphasizes US politics and lifestyle or identity-oriented stories, whereas Chinese news in YACND places relatively more emphasis on national policy, economic development, and social issues. This demonstrates how topic modeling can help reveal cross-lingual differences in media focus and provide an efficient way to summarize large text collections.
Future extensions could include:

- Using more advanced topic models such as BERTopic or neural topic models.
- A multilingual neural topic model such as Top2Vec or BERTopic with multilingual embeddings could reveal deeper semantic alignment beyond bag-of-words assumptions
- Incorporating time information to analyze how topics evolve.
- Extending to more languages or additional news sources.
- Combining topic modeling with sentiment analysis to examine tone as well as content.

ACKNOWLEDGMENTS

I thank Professor Pang for guidance on the DATA 6150 individual project and the authors and maintainers of the HuffPost and YACND datasets for making their data publicly available. I also thank the developers of scikit-learn and Jieba for providing accessible tools for text processing and topic modeling.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT ’19), Vol. 1. Association for Computational Linguistics, Minneapolis, MN, USA, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[2] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL ’19). Association for Computational Linguistics, Florence, Italy, 4996–5001. https://doi.org/10.18653/v1/P19-1493

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS Edition (EMC^2 ’19). Curran Associates, Inc., Vancouver, Canada. arXiv:1910.01108.

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito,

Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (NeurIPS '19). Curran Associates, Inc., Vancouver, Canada, 8024–8035.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (Jan. 2003), 993–1022.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (Oct. 2011), 2825–2830.

[7] HuffPost News Category Dataset. n.d. Kaggle. Retrieved December 2025 from https://www.kaggle.com/datasets/rmisra/news-category-dataset

[8] Ceshine Lee. 2020. Yet Another Chinese News Dataset (YACND). Kaggle. Retrieved December 2025 from https://www.kaggle.com/datasets/ceshine/yet-another-chinese-news-dataset

[9] Sun Junyi. 2025. Jieba: Chinese Text Segmentation. GitHub repository. Retrieved December 2025 from https://github.com/fxsjy/jieba

[10] Google AI Language. 2018. BERT: Pre-trained Models (bert-base-multilingual-cased). Hugging Face model repository. Retrieved December 2025 from https://huggingface.co/google-bert/bert-base-multilingual-cased