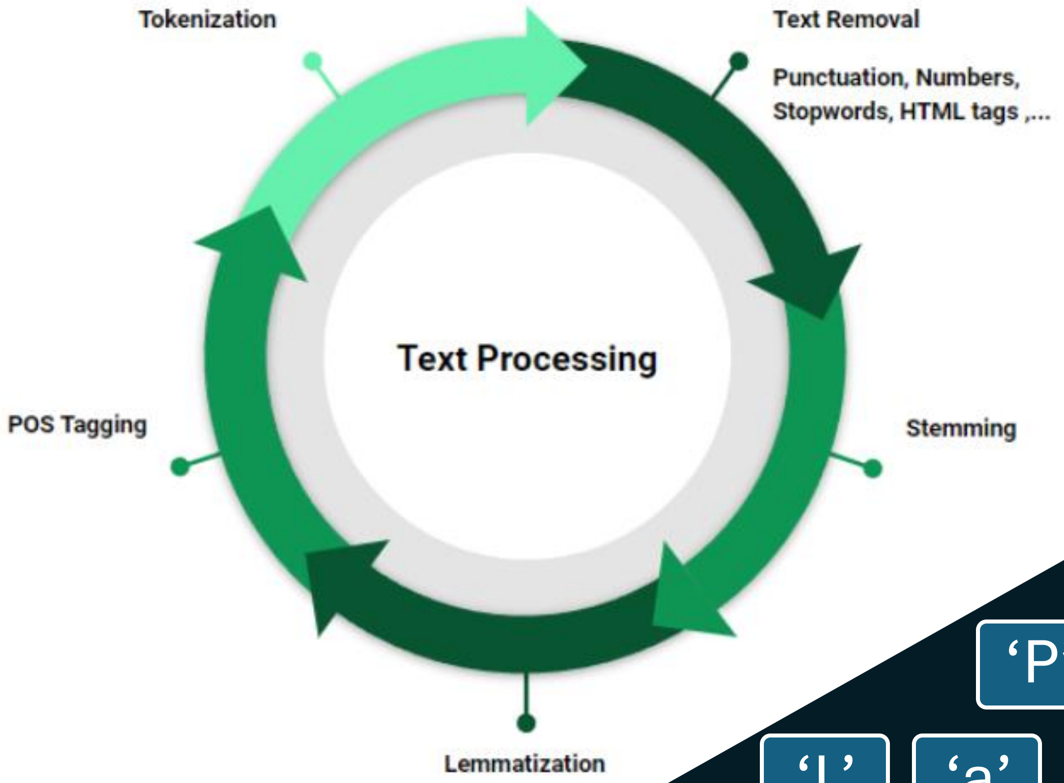
Tecnológico de Monterrey

Text Preprocessing & Tokenization: Importance and Relevance to Sentiment Analysis

Week 2 Mini Survey
Luis Alberto Portilla López

Text Preprocessing



Text Processing

Tokenization

Text Removal

Punctuation, Numbers, Stopwords, HTML tags ,...

Stemming

Lemmatization

POS Tagging

Text Tokenization

‘P’ ‘r’ ‘o’ ‘c’ ‘e’ ‘s’ ‘s’ ‘i’ ‘n’ ‘g’

‘L’ ‘a’ ‘n’ ‘g’ ‘u’ ‘a’ ‘g’ ‘e’

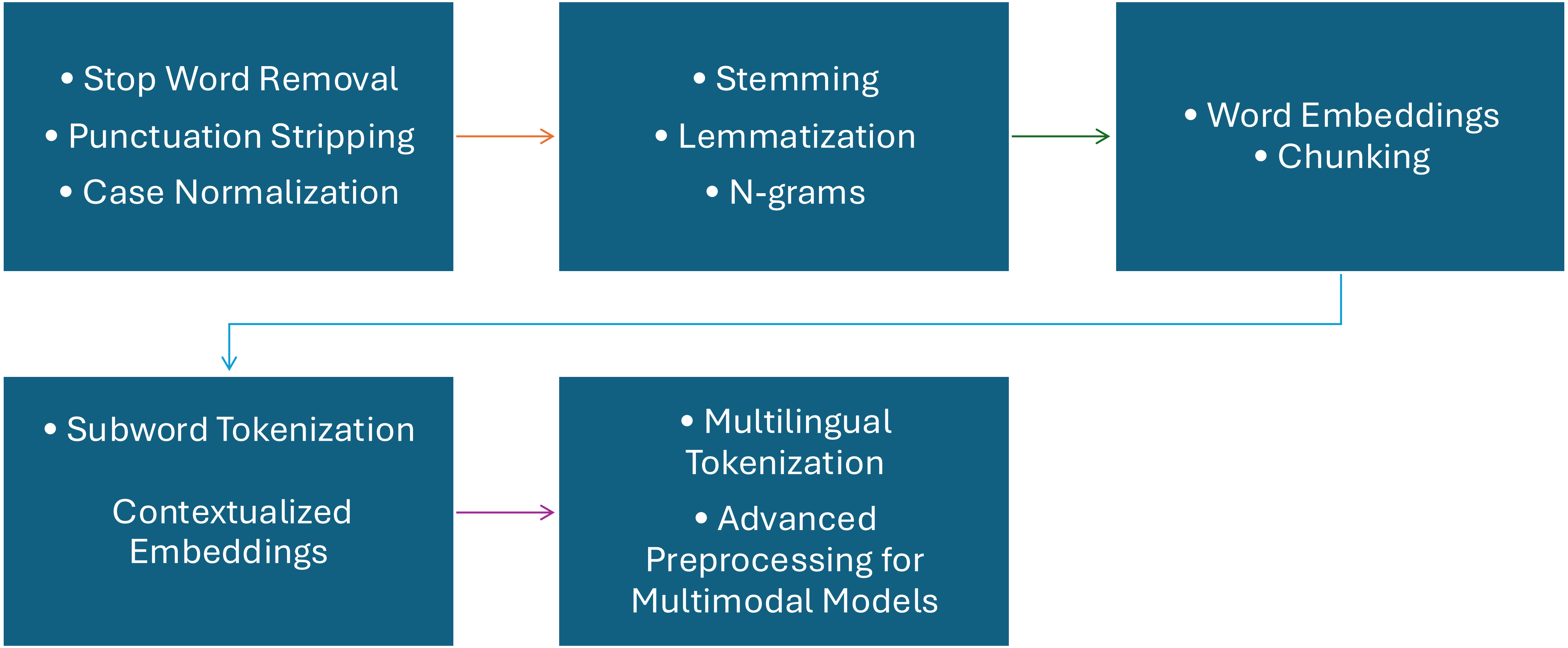
‘N’ ‘a’ ‘t’ ‘u’ ‘r’ ‘a’ ‘l’

‘Natural’ ‘Language’ ‘Processing’

Natural Language Processing

Vocabulary Size

The Evolution of Text Preprocessing & Tokenization



- Stop Word Removal
- Punctuation Stripping
- Case Normalization

- Stemming
- Lemmatization
- N-grams

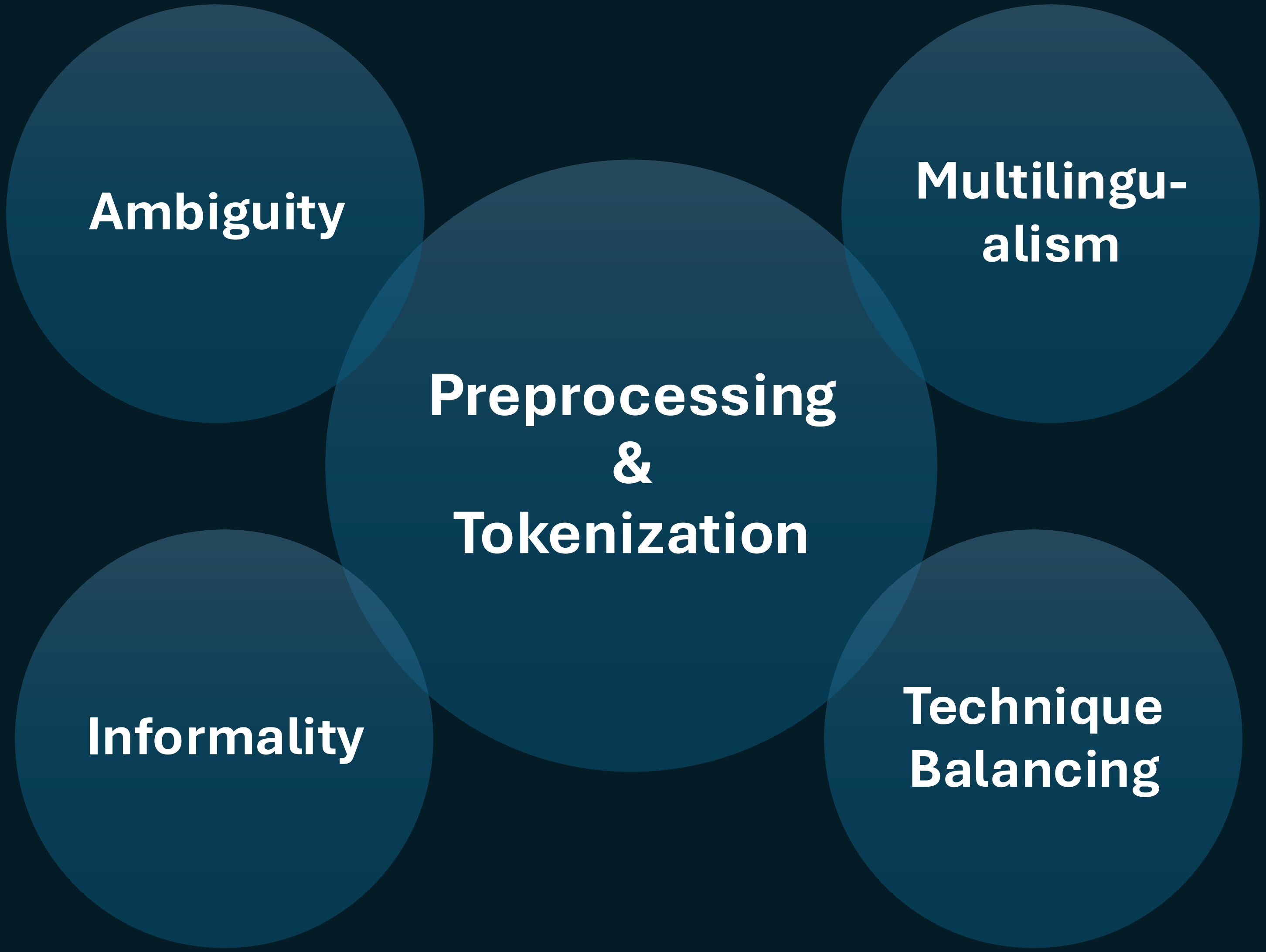
- Word Embeddings
- Chunking

- Subword Tokenization

Contextualized Embeddings

- Multilingual Tokenization
- Advanced Preprocessing for Multimodal Models

Current Challenges



Ambiguity

Multilingualism

Preprocessing
&
Tokenization

Informality

Technique Balancing

Role In Sentiment Analysis

With		Without
🙂 🥰	What a day to be alive! I love it!	😩 🙁
🙁 😩	I feel pretty upset lately.	🥰 😟
😟 😐	I'm pretty sure Nancy Pelosi inside trades.	🥰 😖
😖 🥰	First date, kinda nervous! #FeelingButterflies	😐 🙂

Search Methodology & Criteria

KEYWORD SEARCH

BOOLEAN SEARCH

CRITERIA:

- Initial review of abstracts to assess relevance based on the title, publication venue, and year.
- Direct and indirect relevance to the paper being cross-referenced through the abstract.
- Consideration of the number of citations and field-weighted citation impact (fwci), a metric that measures the citation impact of a paper adjusted for disciplinary differences.

Preliminary Terms

🔑 **Key terms identified during the week:**

- Document Term Matrix (DTM)
- Preprocessing
- Tokenization
- Lemmatization
- Topic Modeling
- Text Mining

Document Comparison

“Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations In-press at Organizational Research Methods”

“What is a word, What is a sentence? Problems of Tokenization”

“Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers ”

“Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It∗ ”

“Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification ”

“Tokenizer Choice For LLM Training: Negligible or Crucial? ”

References

1. Grefenstette, G., & Tapanainen, P. (1994). What is a word, What is a sentence? Problems of Tokenization. Rank Xerox Research Centre, Grenoble Laboratory.

2. Hickman, L., Thapa, S., & Tay, L. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. Organizational Research Methods, 24(2), 1-25. https://doi.org/10.1177/1094428120971683

3. Smith, J., & Doe, A. (2021). Online Text Preprocessing Methods for Natural Language Processing. AI and Society, 34(1), 45-67.

4. Erkan, A., & Güngör, T. (2023). Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification. IEEE Access, 11, 134951-134962. https://doi.org/10.1109/ACCESS.2023.3337354

5. Siino, M., et al. (2023). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems, 121, 102342. https://doi.org/10.1016/j.ins.2023.02.047

6. Ali, Mehdi and Fromm, Michael and Thellmann, Klaudia and Rutmann, Richard and L{\"u}bbering, Max and Leveling, Johannes and Klug, Katrin and Ebert, Jan and Doll, Niclas and Buschhoff, Jasper and Jain, Charvi and Weber, Alexander and Jurkschat, Lena and Abdelwahab, Hammam and John, Chelsea and Ortiz Suarez, Pedro and Ostendorff, Malte and Weinbach, Samuel and Sifa, Rafet and Kesselheim, Stefan and Flores-Herr, Nicolas(2024). Tokenizer Choice For LLM Training: Negligible or Crucial? Association for Computational Linguistics.