MIT-Chile Research Workshop

9:30 - 11:00: Lecture, Universidad de Concepción
From Words to Online Content Moderation
Belén Saldías

In these lecture sessions, students will be introduced to natural language processing and modern out-of-the-box tools and techniques that will allow students to plug & play with many of their language-based datasets in the future. The guiding application and research scenario is content moderation, which we all get exposed to as students. These sessions will open a window to understanding some of the decisions designers and engineers are faced with while architecting such systems. In addition, students will be expected to engage in discussions and hands-on activities designed to further the intended learning outcomes.

Intended Learning Outcomes

- By the end of this session, students will be able to:
- Understand opportunities to develop technology that enables human-centered content moderation.
- Understand text-based sentiment analysis, through interaction with Natural Language Processing (NLP) methods.
- Reflect on ethical considerations that apply to their current research.

Notebooks

1. NLP & Content Analysis
2. Do it yourself

Suggested Readings

Word2Vec and WordEmbeddings

- Bag-of-words model: https://en.wikipedia.org/wiki/Bag-of-words_model
- Tf-idf: https://en.wikipedia.org/wiki/Tf–idf
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/pdf/1301.3781.pdfì—%20ì„œ
- Beyond sentiment analysis:
  Perspective API: Overview, Key Concepts, Scores, Attributes & Language, Model Cards, Training Data, FAQs
  https://developers.perspectiveapi.com/s/about-the-api
  https://www.perspectiveapi.com/
- Human-AI collaboration: Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-24. https://dl.acm.org/doi/abs/10.1145/3359152

Additional Resources

1. Word Embeddings
2. Perspective API