MSc in Computational Software Techniques in Engineering

Noisy Entropy in Languages

Exploring Communication Complexity: A Study of Language Patterns in Social Media and Programming Languages

> Language both written and spoken, exhibits inherent **complexities** that reflect the **unpredictability** and intricate **patterns** of human communication.
> Entropy serves as a quantifiable measure of this **complexity**, capturing the **pattern** and **nuances** within various forms of interaction.

OBJECTIVES

> **Explore Entropy in Language**: Analyze English, French, and Spanish on Twitter as jargon akin to spoken language

> **Analyze Linguistic Chaos in Twitter:** Examine word usage, accents, emojis, and punctuation

> **Investigate Literature Noisy Entropy**: Assess noisy characteristics in literature books

> **Map Noisy Environments**: Focus on contexts as COVID-19 and the Ukraine War

> **Examine Human Emotion**: Uncover emotional patterns and entropy across languages

> **Assess Media Language Styles**: Study news outlets' unique entropy patterns

> **Contrast Personal Communication Tactics**: Compare entropy in public figure tweets

> **Analyze Entropy over Time**: Investigate entropy changes during significant global events

> **Explore Programming Languages Entropy**: Study variables, strings, and numbers in programming languages and their roles

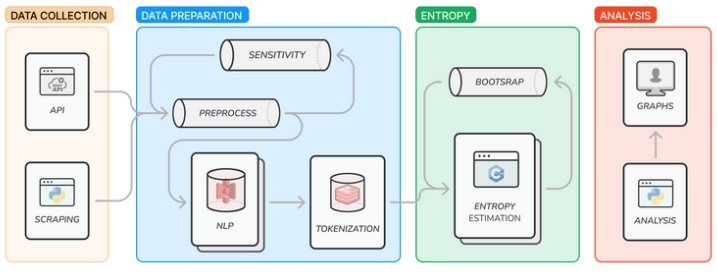
ENTROPY ESTIMATORS

> **Plug-In Estimators: 7 Unique Approaches:**
- Provides a non-parametric approach to entropy estimation
- Work with the statistical distribution of tokens
- Reflect complexity at the token level without considering context

> **Entropy Rate:**
- Measures the average uncertainty per symbol or token
- Leverages *LZ78* compression with a sliding window for efficient estimation
- Captures long-term behavior of information sources

> **Prediction by Partial Matching (PPM):**
- A context-based adaptive statistical data compression technique
- Utilizes adaptive n-gram models with Markov chains
- Offers a flexible approach for pattern recognition and sequence prediction

WORKFLOW



Authors: Siméon Ferez
Contact: simeon.ferez@gmail.com
Supervisors: Dr. Jun Li, Dr Barnes Stuart

KEY METRICS

> **3** Languages (EN, FR, ES)
> **3** Programming Languages (C++, Python, Java)
> **7** Custom Datasets
> **+300** Millions of Tweets Collected
> **~80Gb** of Data processed
> **3** Family of Entropy Estimators
> **7** Plug-In Entropy Estimators
> **8** NLP Analysis Performed
> **2** Methods of Uncertainty Analysis
> **5** Literature Book Examined
> **2** Global Events Studied (Covid-19, Ukraine War)

DATA COLLECTION

> Utilized **Web Scraping** and **APIs** on Twitter
> Leverage **CodeNet** dataset for programming language analysis

STACK

IMPLICATIONS

> **Educational Strategies**
- **Application**: Tailored reading materials
- **Benefit**: Individualized learning, enhanced literacy

> **AI & NLP Techniques**:
- **Application**: Custom preprocessing methods
- **Benefit**: Efficient machine learning models, noise reduction

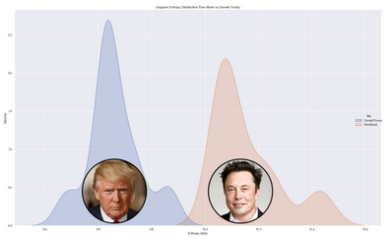
> **Media Literacy & Bias Detection**:
- **Application**: Entropy measurements in journalism.
- **Benefit**: Reader awareness, automated bias detection.

SOME KEY INSIGHTS

> **Entropy in Noisy Literature:**
- Character-Level Insights: Novel estimates of ~4.35 bits, surpassing previous works
- Word-level confirms prior work, and observe a 3 bits shift with context consideration

> **Entropy on Twitter:**
- Character Complexity: Aligns with literature, chaos resides in word formation
- Unigram Entropy: 10-11 bits, highlighted by *hapax legomena* (2 bits above past studies)
- Contextual Shift: A consistent 3-bit shift reveals logical structure despite noise

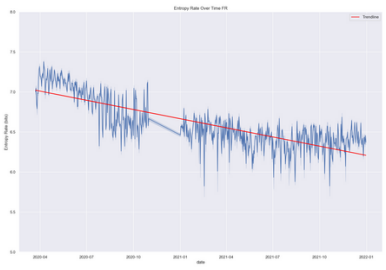
> **Entropy in Personal Communication**:
- Public Figures: Contrast in entropy between Musk (technical diversity) and Trump (political discourse)
- Political Structure: Repeated patterns in politics converge public opinion, low entropy



Entropy Rate comparison between Donald Trump and Elon Musk

> **Cluster Analysis:**
- Symbols and Structure: Accents, emojis increase entropy, punctuation reduces it, defying traditional beliefs
- Emotional Patterns: Universal low entropy in inherent emotions like love & fear
- Media Bias: News outlet styles reveal unique entropy, hinting at bias detection capabilities

> **Entropy over Time**
- COVID-19, Ukraine War show unigram entropy convergence, highlights societal response, unity
- Language Evolution: Decline in entropy signifies societal adaptation, convergence in viewpoints



Entropy Rate trend over time of the Covid-19 French dataset

> **Entropy in Programming Languages**
- Universality: Common features in C++, Python, Java
- Predictable Structure: Low Entropy Rate highlights codified patterns and conventions
- Source of Noise: Variable and function naming contribute most to complexity