# RESEARCH STAY WEEK 2, TEXT PREPROCESSING

Daniel Cajas A01708637

# CONTEXT

## Why talk about text preprocessing?

Even before the hype around new large language models (LLMS), like GPT4, there has been a lack of formal studies around the effects that standard industry techniques have in the behavior and accuracy of those models.  Here we try to understand the effects that have been found and highlight their importance.

The techniques we are going to look into are:

- **Punctuation removal and interpretation**
- **Number removal**
- **Lowercasing**
- **Stemming**
- **Stop word removal**
- **N-grams**
- **Infrequently used terms removal**

# SEARCH METHODOLOGY

The material given was a useful introduction to how text is preprocessed but it is a little outdated. Mainly what i felt it lacked was how it relates to state of the art NLP techniques, specially transformers. That is why i looked for material similar to “Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations”.

Through google scholar I got to this promising text. Marco Siino, Ilenia Tinnirello, Marco La Cascia, Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers Information Systems, Volume 121, 2024, 102342, ISSN 306-4379, https://doi.org/10.1016/j.is.2023.102342.

# USEFUL ACRONYMS

**Table 1**
Acronyms for the preprocessing techniques and real case examples, raw and preprocessed.

| Acronym | Technique | Raw | Preprocessed |
|---|---|---|---|
| DON | Do Nothing | "Like a Rolling Stone" | "Like a Rolling Stone" |
| RNS | Replace Noise | *"@Obama tells #metoo! bit.ly/–"* | *"USER tells HASHTAG! URL"* |
| RSA | Replace Slang/Abbreviations | *"omg you are so nice!"* | *"Oh my God you are so nice!"* |
| RCT | Replace Contraction | *"I don't like butterflies."* | *"I do not like butterflies."* |
| RRP | Remove Repeated Punctuation | *"I like her!!!"* | *"I like her multiExclamation"* |
| RPT | Removing Punctuation | *"You. are. cool."* | *"You are cool"* |
| RNB | Remove Numbers | *"You are gr8."* | *"You are gr."* |
| LOW | Lowercasing | *"You Rock! YEAH!"* | *"you rock! yeah!"* |
| RSW | Remove Stop Words | *"This is nice"* | *"is nice"* |
| SCO | Spelling Correction | *"1lenia is so kind!"* | *"Ilenia is so kind!"* |
| POS | Part-of-Speech Tagging | *"Kim likes you"* | *"Kim (PN) likes (VB) you (N)"* |
| LEM | Lemmatization | *"I am going to shopping"* | *"I be go to shop"* |
| STM | Stemming | *"Girl's shirt with different colors"* | *"Girl shirt with differ color"* |
| ECR | Remove Elongation | *"You are cooool!"* | *"You are cool!"* |
| EMO | Emoticon Handling | *":)"* | *"happy"* |
| NEG | Negation Handling | *"I am not happy today!"* | *"I am sad today!"* |
| WSG | Word Segmentation | *"#sometrendingtopic"* | *"some+trending+topic"* |

(Marco Siino, 2024)

# COMPARISON

The exact results can not be directly compared but the the general conclusion, that preprocessing has a significant impact in model performance, is the same. The better preprocessing steps are not even determined by the model being used as different corpus using the same preprocessing techniques have varying results even on the same models.

While this general conclusions are the same, how this conclusions were reached was different. Marco focuses less on the inherent structure and nature of the corpus data and more on comparing the models themselves distinguishing between the “new” deep learning models and more traditional non deep learning models. This gives a new, useful perspective as pretrained transformers have gained popularity.

Something interesting to point out is that deep learning models did best with either no preprocessing done or just stop word removal. Being pre trained, they seem to depend more on the context given by elements that would be removed with other techniques.