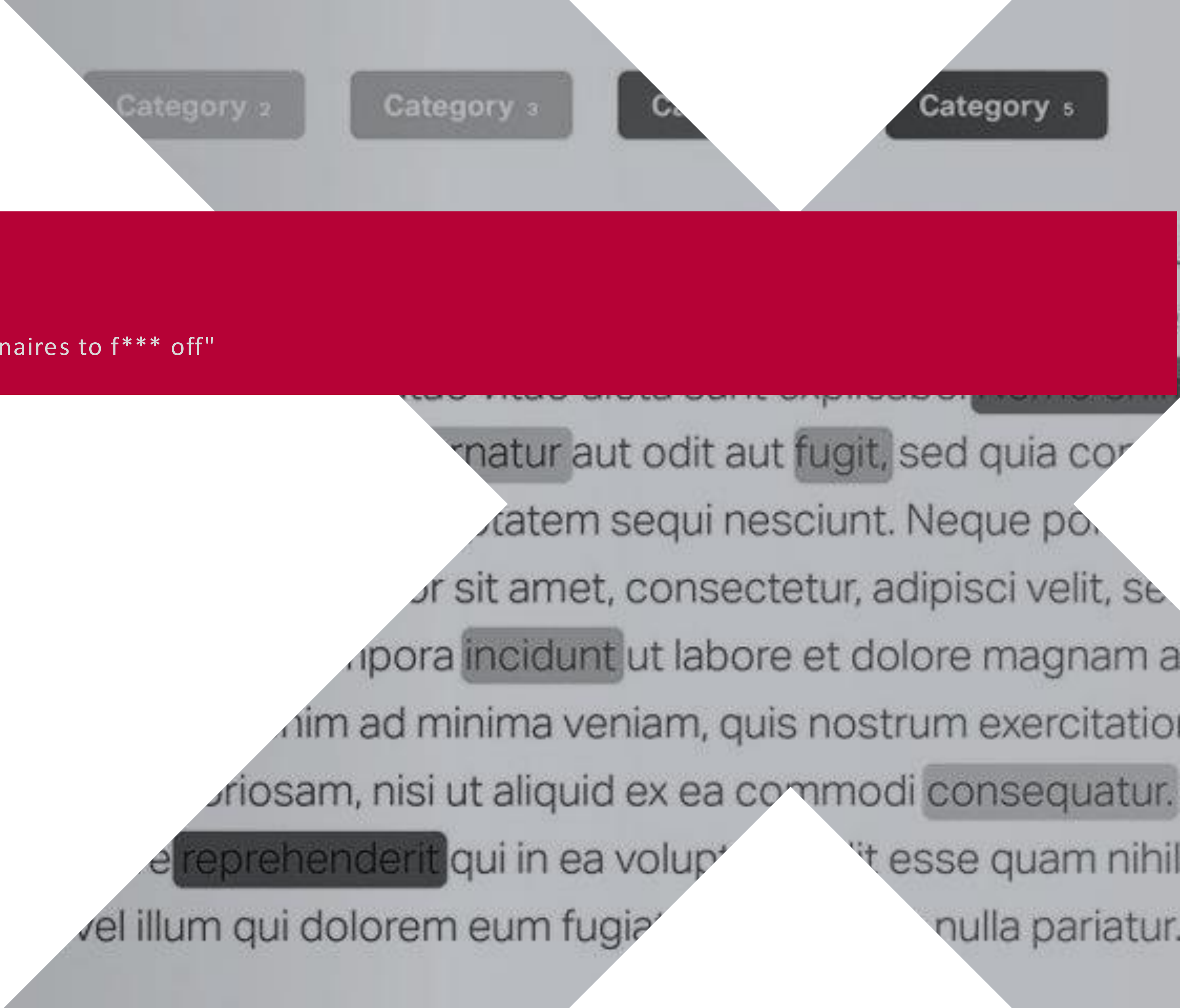
NLP FAKE NEWS

"bernie sanders just told republicans and billionaires to f*** off"



natur aut odit aut fugit, sed quia co

tatem sequi nesciunt. Neque po

r sit amet, consectetur, adipisci velit, se

npora incidunt ut labore et dolore magnam a

nim ad minima veniam, quis nostrum exercitatio

riosam, nisi ut aliquid ex ea commodi consequatur.

e reprehenderit qui in ea volup it esse quam nihil

vel illum qui dolorem eum fugia nulla pariatur.

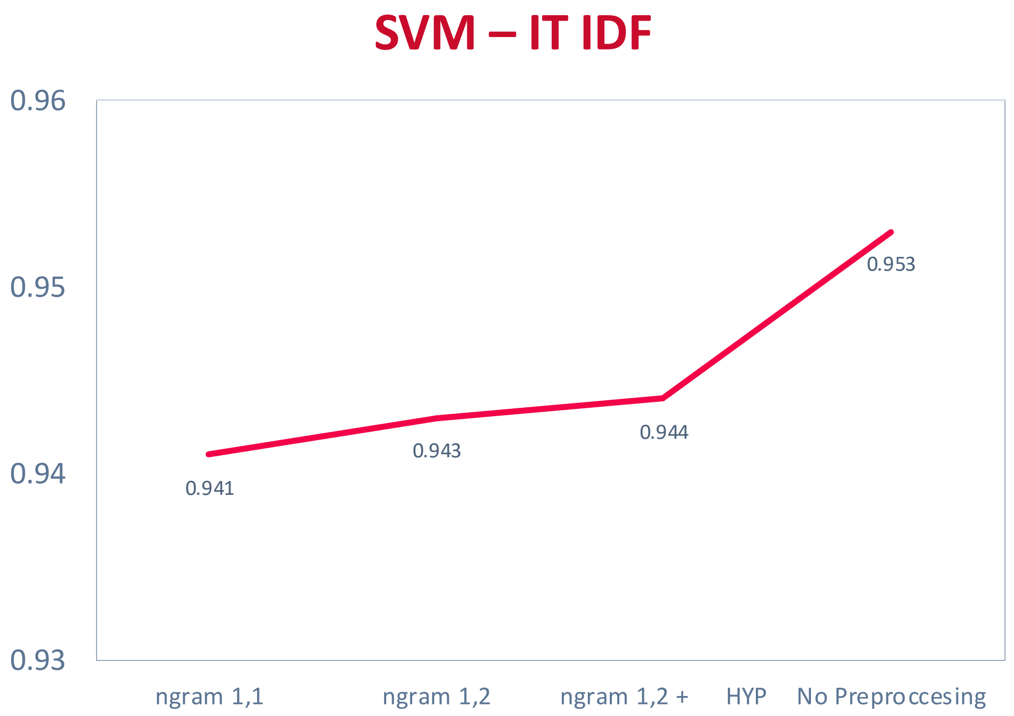
GROUP 3:

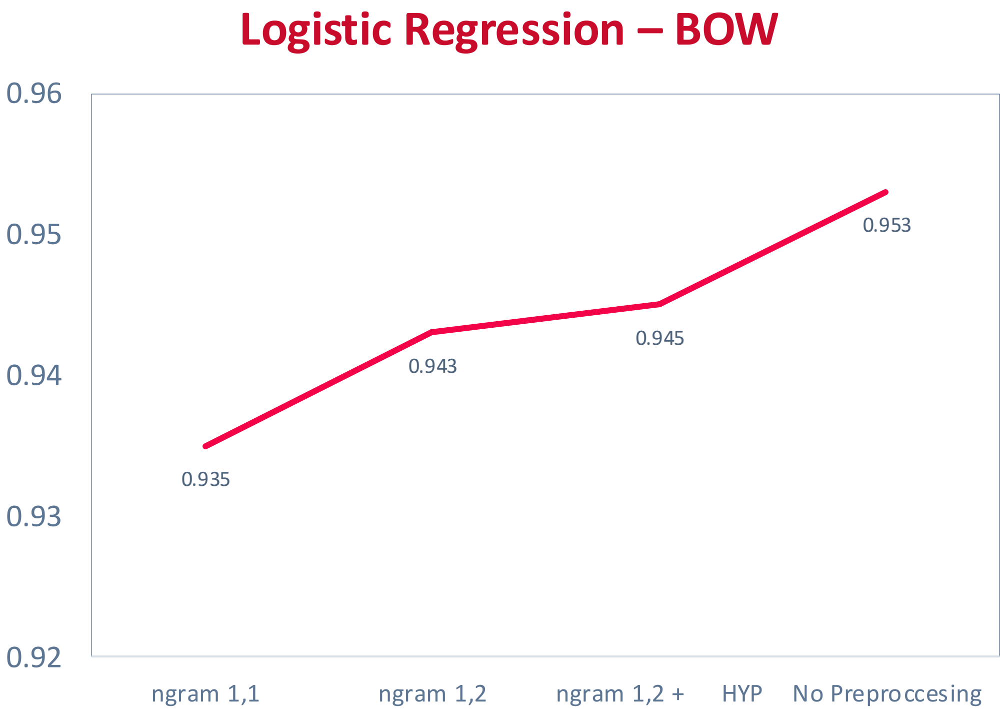
Mo

Ewa

Diego A.

1.- EXECUTIVE SUMMARY

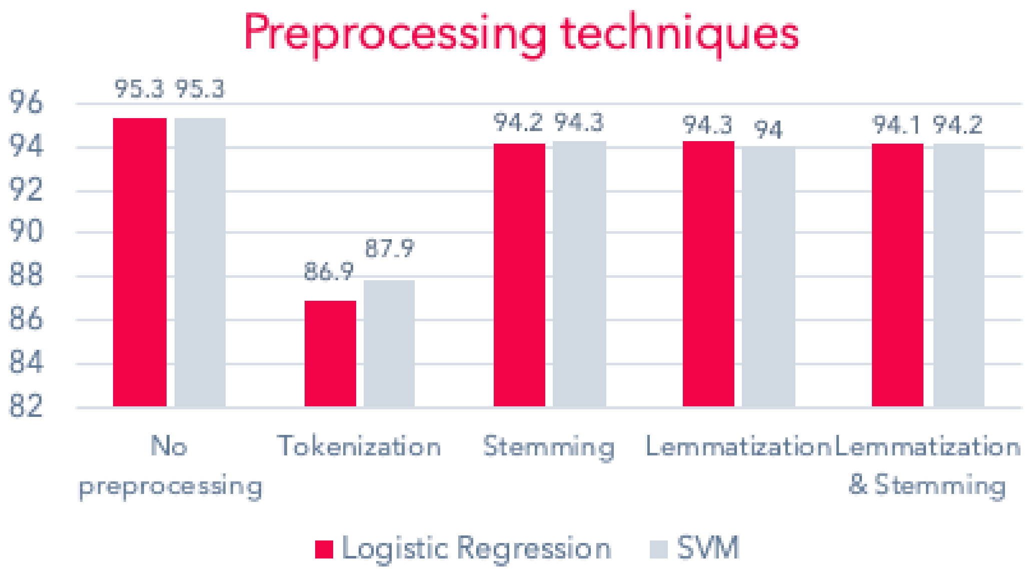




- **Used Models:** Naïve Bayes, Log Reg, Decision Tree, Random Forest, KNN & SG Boost
- **Best Performance:** Our best models achieved an **accuracy rate** of **0.953.** Yet, **Log Reg.** performed much faster
- **SVM** performed slightly better on the F1 score, recall and cm but only with IT IDF; the BoW approach showed strong underperformance
- **Naïve Bayes == Log.Reg** (almost).

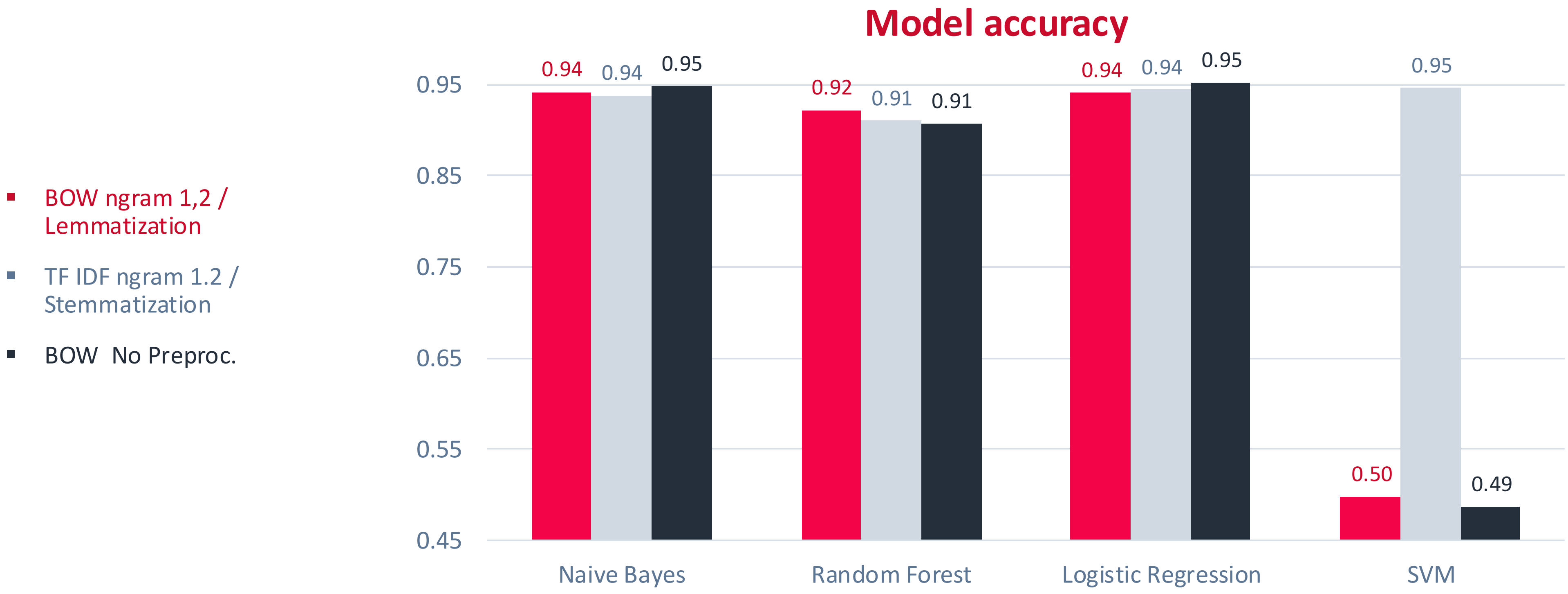
2.- METHODS (PREPROCESSING)

- **Approach:** We tested the accuracy of our models against different versions of our input data
- **Techniques:** Remove punctuation, numbers and special characters. We removed stopwords; tokenized, stemmed and lemmatized the data.
- **Surprise:** No preprocessing yielded the best results
- This indicates a **loss of contextual** information through preprocessing



3.- MODELS

- **Model development strategy:** Initially we tested a broad range of classification models on the first dataset yielding bad results. The performance increased significantly when changing to the Fake news dataset
- **Process:** tested all different models with different ngrams changing preprocessing values an evaluate them using accuracy, report, and confusion matrix. We implemented cross validation with 10 folds to detect overfitting on our best Log-Reg Model



4.- TAKEAWAYS

- **Why not using preprocessing working better?**
  - **Loss of Signal:** We may inadvertently remove or simplify words that carry subtle but important contextual meaning
  - **Dataset characteristics:** In fake news detection, seemingly irrelevant words might contribute to identifying nuances between real and fake news. i.e: swear words, or even symbols such as HTs more used in fake news.
  - **Model dependence:** Simpler models might be better suited for raw data since they thrive on simple word frequency counts
- **Why ngram range 1,2 is performing better?**
  - Adds more context by considering word pairs, often improving performance by capturing relationships between words and also captures individual word frequencies

And this GIF works here and in graphic design 🙃

Thank you !