Process Summary

1. **Segmentation**
   - We **split** each text into smaller chunks (segments) based on punctuation (e.g., splitting on `.` or another delimiter).
   - This ensures shorter pieces are easier to classify accurately.
2. **Classification with Smoothness**
   - Each segment is **classified** using a model that incorporates a smoothness assumption. This means the model discourages frequent switching between labels (e.g., “spam” vs. “story”), helping maintain more coherent classifications across adjacent segments.
3. **Purging Spam, Retaining Stories**
   - Segments labeled as **spam** are removed (“purged”).
   - Segments labeled as **stories** are kept, preserving only the text deemed relevant.

Performance

- **Speed**: ~3 minutes per 1,000 data samples.
- **Scale Estimate**: For 100GB of data, total processing time is approximately **25 hours**.

Caveats & Observations

- **Merged Chunks**: Some text chunks ended up “stuck” together during scraping, making it harder to separate them correctly. There’s little we can do about it other than rescrape.
- **Foreign Language Bias**: Certain foreign-language segments (e.g., Vietnamese) may appear “spammy” to the classifier and can be incorrectly purged in extreme cases.
- **Possible False Purges**: Due to these biases and merged chunks, a small portion of valid text might occasionally be labeled as spam and removed.

Further Experimentation

- **Play with different “prompts”:** the zero shot classifier allows us to describe in our oun words the “labels” of what we want to keep, and what we want to get rid of
- **Work with lighter models**

Examples

Some good text in vietnamese was unrightfully purged:

Your cart is empty!WE ARE HELIOS!Unique design, handcrafted, each piece is a separate story
Not only accompanying men on their journey of discovery, expressing their personal style, Helios always strives to spread the passion and love for works made by Vietnamese people, and at the same time bring those works to the world in a strong and proud way
QUALITY Different from many brands on the market, we are inspired by every aspect of life, combining them with bold thinking to create unique products, with a strong and cool style
Devoted to the creations created by pure Vietnamese hands, Helios wishes to affirm their value to the international community, bringing the item to the world in a strong and proud way
INSPIRATIONExploiting aspects surrounding men, Helios seeks an iconic image with a strong connection to their story
3D MODELFrom hand-drawn sketches, the digital design team creates a 3D model with the help of graphics software
Using high-quality S925 silver, melting it and pouring it into a mold to create a silver tree, each branch corresponding to a product
FINISHING PRODUCTIONWhen the production process is completed, the quality department inspects and evaluates each product, ensuring both the appearance and the actual experience when using it

Some bad text “hides” inside good text, (kept):

Our Story Free shipping on all orders over £95My name is Paul and after years of living in the busy finance environment, I decided it was time to make a change