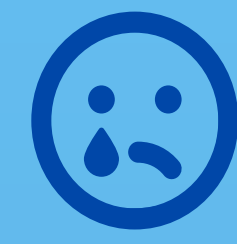
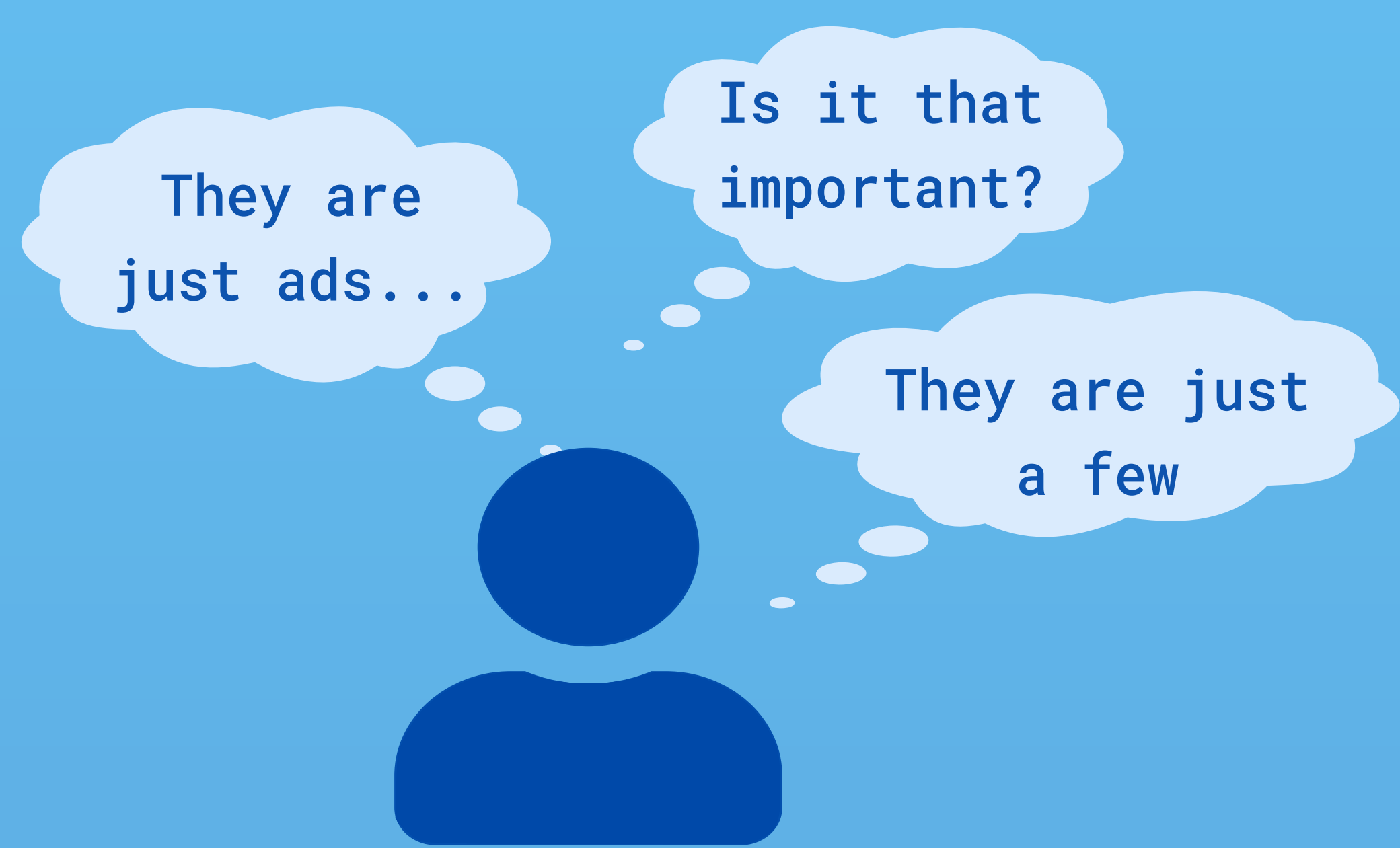




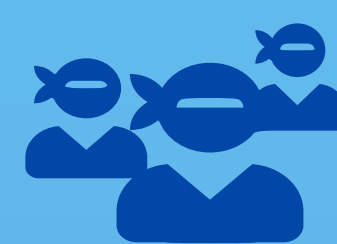
SPAM DETECTION USING NEURAL NETWORKS

What is spam?

Undesired or unsolicited content. In the scope of a digital repository it means "all kinds of content that are not meant to be present in said repository, and therefore, shown to the users".



Bad UX due to unwanted content - unhappy users.



Vast amounts: social networks are fighting the problem and repositories should be no exception.



Growing fast! 1,000% in the last year. 1 in 5 submissions of 2020 were spam.



Monetary costs: storage of content and files, network bandwidth to transmit it, legal consequences...

Challenges

Big difference between the amount of spam and non-spam content (x45 times).

Text content is very similar. Not many distinct clusters.

Text length varies from a few words to 35,000.

Multiple languages. More than 30 different ones, although ~60% is English.

User related data not available.



Actions

Undersample the majority class (reduce non-spam records).

Try different models and neurons based on literature.

Study text length distribution and reduce it to a maximum of 5,000.

Compare weighted accuracy between English-only content and all languages (not significant).



Future work

Mix under and oversampling, or more complex methods such as cluster sampling.

Perform a deeper clustering analysis and find more useful features.

Perform an in depth study of text length distributions between spam and non-spam content.

Test language agnostic models.

Find ways to use user related data while respecting data privacy policies.

Outcome

Tested several neural network based models from literature.

Investigated in depth a mixed model, with both convolutional (CNN) and recurrent (LSTM) layers.

Used embedding layers to produce vectorized representations of the text.

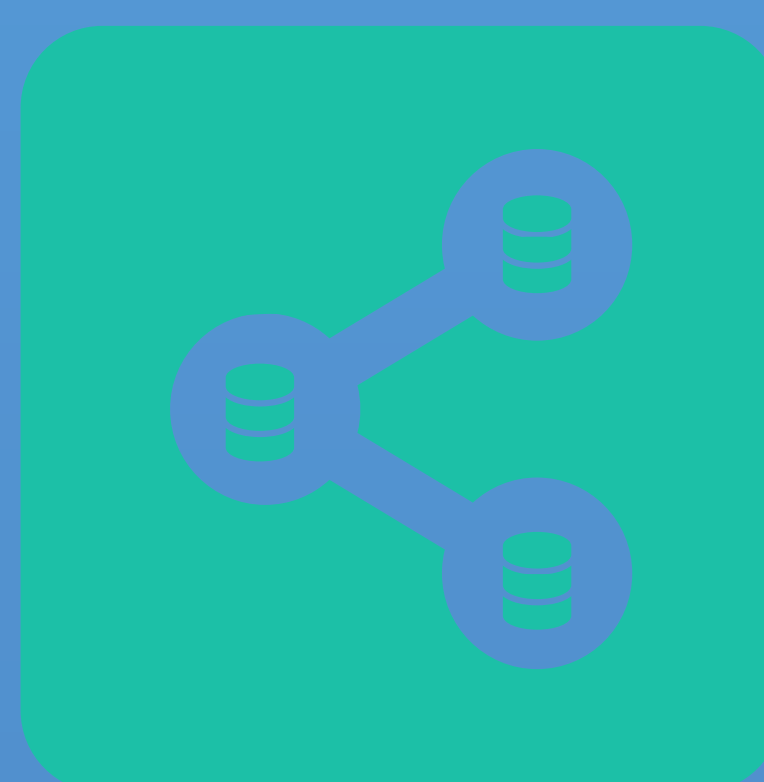
Achieved a ~90% precision on the testing dataset.

High amount of false positives due to a low amount of spam in the training dataset.

The future could be open!



Do you suffer from spam?



BYOS (Bring Your Own Spam)

Create a common spam dataset and share models



How have you tried to solve it?



Visit Zenodo
zenodo.org



Follow us on Twitter
[@zenodo_org](https://twitter.com/zenodo_org)



Contact us
zenodo.org/support

