January 2020

Insult detection and filtering

Christina Abboud

Jyrki Alakuijala

Insult detection and filtering

ABSTRACT

Website publishers, search engines, social media service providers, and other online entities need to comply with regulation that requires removal of content that is deemed to be insulting to an individual. This disclosure describes techniques to match and remove text that is similar to insult strings that mandated for removal. One-way hashes of insult strings are used to enable removal of insulting content without exposing the name of the subject of the insult string. Name canonicalization is performed and a semantic fingerprint is obtained using a one-way function.

KEYWORDS

- Insult detection
- Semantic matching
- Semantic fingerprint
- Content filtering
- Hash function
- One way hash

BACKGROUND

Website publishers, search engines, social media service providers, and other online entities need to comply with regulation that requires removal of content that is deemed to be insulting to an individual. Such content is termed as an unwanted insult. The mandate can automatically extend to a larger class of content, e.g., insults that are similar to the unwanted insult to be removed, and to multiple sources of online content. While a list of insults or

matching patterns can be used to filter online content, storing or transmitting such a list can expose such insults to any listener or an entity that filters content.

DESCRIPTION

This disclosure describes techniques to match and remove text that is similar to insult strings that mandated for removal. One-way hashes of insult strings are used to enable removal of insulting content without exposing the name of the subject of the insult string. Name canonicalization is performed and a semantic fingerprint is obtained using a one-way function.

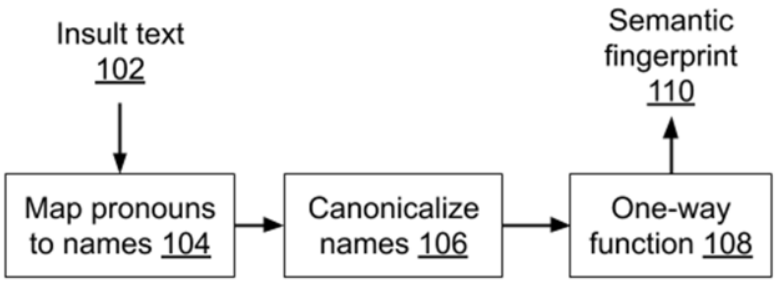

Fig. 1: Insult detection and filtering

Fig. 1 illustrates insult detection and filtering, per the techniques of this disclosure. Insult text (102) is obtained, e.g., from a court mandate or from a mandate by another regulatory body. The insult text is automatically analyzed to map pronouns to names (104). Names or identities that appear in the insult text are canonicalized (106), e.g., via transliteration. A semantic fingerprint or hash (110) is obtained using a one-way function (108).

The use of a one-way function enables a larger class of content, e.g., insults similar to the unwanted insult or translation of the insult to languages other than the original language of the insult, to be identified. The one-way function can be based on the latent space description of a language model, e.g., the one-way function can rotate the latent space of a language to align with the latent space of another language. The one-way function can also be based upon manually-

written rules, e.g., rules that declare the semantic equivalency of insult-string A to insult-string B. The one-way function enables content filtering without wide distribution of unwanted insults.

In this manner, an array of hashes or semantic fingerprints is used to distribute a list of unwanted insults, limiting knowledge of the actual content of the unwanted insult, while enabling removal of matching content. The list of hashes does not expose names or corresponding insult text.

CONCLUSION

This disclosure describes techniques to match and remove text that is similar to insult strings that mandated for removal. One-way hashes of insult strings are used to enable removal of insulting content without exposing the name of the subject of the insult string. Name canonicalization is performed and a semantic fingerprint is obtained using a one-way function.

REFERENCES

[1] Sarah Downey, “Online, you’re guilty even after being proven innocent” https://www.abine.com/blog/2017/online-guilty-before-proven-innocent/, accessed Dec. 29, 2019.