ABSTRACT

Automatic detection of hate speech in social media

Claudia Zaghi

Supervisors: Malvina Nissim - Alber Gatt

Hate speech is "the use of aggressive, hatred or offensive language, targeting a specific group of people sharing a common trait: their gender, ethnic group, race, religion, sexual orientation, or disability" (Mish and Morse, Mish and Morse).

As the phenomenon is widely spreading online (Gagliardone et al., 2015), social networks and websites have introduced a progressively stricter code of conduct and regularly removed offensive content flagged by users (Bleich, 2014). However, the volume of data in social media makes it challenging to supervise the published content across platforms.

This research focuses on hate speech in Italian, aiming to automatically detect hateful content based on data scraped from different social media sites. Using the technique of distant supervision (Go et al., 2009), we automatically developed labeled datasets for machine learning experiments as well as hate-polarized word embeddings.

We tackled the challenge of hate speech detection by training a simple binary classifier, characterized by a Linear Support Vector Classification (SVC) model and n-grams features. We compared the performance of the classifier when trained and tested over manually and automatically annotated datasets, and resources containing hatred against a single target versus multiple targets.

The results of the study highlighted the effectiveness of manually labeled data (80% vs. 45% in F-1 score) and the versatility of distantly supervis data, as sections of automatically labeled data can be used to enrich sm manually labeled datasets. The polarized word embeddings proved to more predictive than off-the-shelf dense vectors (81% vs. 79% in F-1 sco Additionally, the experiments showed that the language of *haters* is very similar across targets.