Profiling Hate Speech Spreaders on Twitter

Sharon Yang and Dr. Sole Pera (Faculty Advisor)
Department of Computer Science

BOISE STATE UNIVERSITY

BACKGROUND

Hate speech is defined as any public communication that depreciates a person or a group by expressing hate or encouraging violence[1].
From the identification of the profiles of hate propagators, it is possible to avoid the spread of hate speech and keep social networks healthier.
In this study, we focused on **Twitter**.

PURPOSE

Simply analyzing words in tweets is a good starting point to identify hate speech and people who spread hate speech. However, we believe there is value in considering other expressions that are commonly seen in tweets.
The purpose of this study was to explore a variety of expressions and unveil a set of common patterns that could lead to identifying user profiles that promote hate speech on social media (Twitter).

DATA

We examined data for **200 Twitter users**.
Each user has a set of tweets and the corresponding label (hate speech spreader or not)[2] .

METHODS

Inspired by the work presented in [2], we investigated features beyond words that can facilitate hate spreader detection.
Following the experimental framework in [2], we first considered **expressions** often seen in tweets. We then outlined **features** that enable us to capture these expressions from multiple perspectives.
We treated hate speech spreader detection as a **classification** problem and therefore, we relied on well-known **classifiers** for hate speech spreader prediction.
Finally, we **evaluated** the impact of the proposed features using traditional classification metrics.

EXPRESSIONS

- Fully capitalized words
  e.g. SCHOOL
- Emoji
  e.g. ❤️❤️
- Special Character
  e.g.!@#$%^&*?_=+-*[]<>()/\|{}~;:'``…”...“
- Hashtag
  e.g. #breastcancer
- Mention
  e.g. @FSMagazine
- URL
  e.g. http://ow.ly/rydSa
- Retweet
  e.g. RT
- Misspelling
  e.g. retreival

FEATURES

- Frequency of occurrence of a given expression. (Applied to all expressions)
- Maximum and Average length of a sequence on character level pertaining a particular expression. (Applied to expressions: fully capitalized words, emoji)
- Maximum and Average length of a sequence on word level pertaining a particular expression. (Applied to all expressions)

CLASSIFIERS

In this study, we used two well-known classifiers following the experimentation from [2]:

- Support Vector Machines (**SVM**)
- Random Forest (**RF**)

RESULTS

Expression	Classifier	Accuracy
Fully capitalized words	SVM	0.425
	RF	0.575
Emoji	SVM	0.475
	RF	0.475
Special Character	SVM	0.475
	RF	0.525
Hashtag	SVM	0.475
	RF	0.375
Mention	SVM	0.475
	RF	0.525

Expression	Classifier	Accuracy
URL	SVM	0.475
	RF	0.400
Retweet	SVM	0.475
	RF	0.650
Misspelling	SVM	0.425
	RF	0.550
ALL	SVM	0.425
	RF	0.475

Summary of the results obtained using an 80/20 data split for training and testing.

CONCLUSION

- When considering features associated with the studied expressions, the accuracy rates are within the range of [0.4, 0.675].
- No expression seems to dominate the outcome.
- In general, increasing the number of features does not always increase the overall accuracy of the classifier.
- In the future, we would like to combine features based on words as well as the expressions we examined to attempt to further increase hate speech spreader detection.

References
[1] Wikipedia contributors. "Hate speech." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 26 Nov. 2021. Web. 29 Nov. 2021.
[2] Claudio Moisés Valiense de Andrade and Marcos André Gonçalves. Profiling Hate Speech Spreaders on Twitter: Exploiting Textual Analysis of Tweets and Combinations of Multiple Textual Representations. In Guglielmo Faggioli et al., editors, CLEF 2021 Labs and Workshops, Notebook Papers, September 2021. CEUR-WS.org.