HaMor
at
the Profiling Hate Speech Spreaders on Twitter

Notebook for PAN
at CLEF 2021

Mirko Lai[1], Marco Antonio Stranisci[1], Cristina Bosco[1], Rossana Damiano[1] and Viviana Patti[1]
1. Università degli Studi di Torino, Italy

Profiling Hate Speech Spreaders on Twitter

HATER

given a Twitter feed of 200 messages, determines whether its author spreads hatred content or not

NOT HATER

Spanish
English

HS MV NE CB

Our Proposal

we proposed 4 types of features:

1. Hate Speech Detection (**HS**)
   a. models-based
   b. lexica-based
2. Moral Foundation Theory (**MV**)
3. Named Entity Recognition of Hate Speech target (**NE**)
4. Communicative Behavior (**CB**)

HS

1. Hate Speech Detection: models-based

HatEval is a dataset for hate speech detection against **immigrants** and **women** in **Spanish** and **English** tweets (Task 5 of the SemEval-2019 workshop)

we trained 3 different models for counting

- for each user -

the number of hateful tweets predicted

[1] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, **SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter**, in: Proceedings of the 13th International Workshop on SemanticEvaluation, Association for Computational Linguistics, 2019, pp. 54–63.

HS

1. Hate Speech Detection: models-based

- SemEvalSVM (**SESVM**): a linear SVM trained using a text 1-3 grams bag-of-words

- Atalaya (**ATA**) [1]: linear-kernel SVM trained on a text representation composed of bag-of-words, bag-of-characters and tweet embeddings, computed from fastText wordvectors (**Spanish**)

- Fermi (**FER**) [2]: RBF kernel trained on tweet embeddings from Universal Sentence Encoder (**English**)

[1] J. M. Pérez, F. M. Luque, **Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification**, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 64–69.
[2] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, V. Varma, **FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter**, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 70–74.

HS

1. Hate Speech Detection: lexica-based

- HurtLex (**HL**): a lexicon of offensive, aggressive, and hateful words in over 50 languages (including English and Spanish). [1]

- NoSwearing [www.noswearing.com] (**NoS**): a list of English swear, bad, and curse words. Spanish translation by Pamungkas et. al [2]

- the Racial Slur Database [http://www.rsdb.org/full] (**RSdb**): a list of English words that could be used against someone: a specific race, sexuality, gender etc. Spanish translation by Babelnet's API

[1] E. Bassignana, V. Basile, V. Patti, **Hurtlex: A multilingual lexicon of words to hurt**, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6
[2] E. W. Pamungkas, A. T. Cignarella, V. Basile, V. Patti, et al., **14-exlab@ unito for ami atibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets**, in: 3rd Workshop on Evaluation of Human Language Technologies for IberianLanguages, IberEval 2018, volume 2150, CEUR-WS, 2018, pp. 234–241.

MV

2. Moral Foundation Theory [1]

The main five foundations

care/harm

fairness/cheating

loyalty/betrayal

authority/subversion

purity/degradation

Could the foundations **loyalty/betrayal** and **authority/subversion** be related to **hatred** contents online?

[1] J. Haidt, C. Joseph, et al., **The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules**, The innate mind 3 (2007) 367–391.

MV

2. Moral Foundation Theory

- extended Moral Foundations Dictionary (**eMFD**) [1]

  is a dictionary of English terms categorized by a specific moral foundation

  the mean, the standard deviation, and the total amount of terms occurring in author feed

  four categories **loyalty/betrayal** and **authority/subversion**

  translated in Spanish by Babelnet's API

[1] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, R. Weber, **The extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text**, Behavior Research Methods 53 (2021) 232–246

MV

2. Moral Foundation Theory

- Moral Foundations Twitter Corpus (**MFTC**) [1]

is a collection of 35,000 English tweets annotated for their moral domains

transfer learning

convert the original multi-label annotation schema in a binary-label one

loyalty, betrayal, authority or subversion => **HS (true)**

while the other => **HS (false)**

[1] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy,M. Atari, Z. Kamel, M. Mendlen, et al., **Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment**, Social Psychological and Personality Science 11(2020) 1057–1071

MV

2. Moral Foundation Theory

- Moral Foundations Twitter Corpus (**MFTC**) [1]

- for each user -

we computed the number of potential hateful tweets

predicted by a linear SVM trained using a text 1-3 grams bag-of-words representation of the *Moral Foundations Twitter Corpus*

<u>Available only for English language</u>

NE

3. Named Entity Recognition of Hate Speech target

the mention of a person belonging to a group vulnerable to discrimination

could be a **signal of hatred contents**

could help in discriminating between what is HS and what is not

NE

3. Named Entity Recognition of Hate Speech target

1. PERSON entities detected by the **transition-based named entity recognition** component of **spaCy**
2. The retrieved entities have been searched on **Wikipedia** through the **Opensearch API**
3. we recover the entries of the category box of each Wikipedia page
4. we manually selected the categories that could be **subject to discrimination**

RT #USER#: Day 39: Biden to Reopen A Detention Facility **Kamala Harris** Protested Against. #URL#

```
['Kamala', 'Kamala Harris', 'Kamal Haasan',
'Kamala (wrestler)', 'Kamala Khan', 'Kamala Surayya',
'Kamala Harris 2020 presidential campaign',
'Kamaladevi Chattopadhyay', 'Kamala Mills fire',
'Kamalani Dung']
```

Categories: Kamala Harris | 1964 births | 21st-century American memoirists | 21st-century American politicians | 21st-century American women politicians | 21st-century American women writers | African-American candidates for President of the United States | African-American candidates for Vice President of the United States | African-American members of the Cabinet of the United States | African-American memoirists | African-American people in California politics | African-American United States senators | African-American women in politics | African-American women lawyers | American people of Indian Tamil descent | American politicians of Indian descent | American politicians of Jamaican descent | American prosecutors | American women lawyers | American women memoirists | Asian-American members of the Cabinet of the United States | Asian-American United States senators | Baptists from California | Women vice presidents | Writers from Oakland, California

NE

3. Named Entity Recognition of Hate Speech target

we obtained two gazetteers of potential HS targets
75,890 entities for English, and 31,235 for Spanish

```
{Margaret Skirving Gibb : Scottish feminists,
 Melih Abdulhayoğlu : Turkish emigrants to the USA,
 James Adomian : LGBT people from Nebraska [...]}
```

the feature expresses:

- the total number of potential HS targets mentioned in author feed, the mean, the standard deviation
- the ratio between the number of HS target
- the total amount of HS targets mentioned by the author in her/his feed

CB

4. Communicative Behaviour

the **structure** of the tweet and to the **user's style**

- Bag of Words (**BoW**): binary 1-3 grams of all author's feed.
- Bag of Emojis (**BoE**): binary 1-2 grams of all author's feed only including emojis

CB

4. Communicative Behaviour

- Uppercase Words (**UpW**):
  the amount of words starting with a capital letter and the number of words containing at least two uppercase characters.

- Punctuation Marks (**PM**):
  the frequency of exclamation marks, question marks, periods, commas, semicolons, and finally the sum of all the punctuation marks mentioned before.

- Length (**Len**): 3 different features were considered to build a vector: number of words, number of characters, and the average of the length of the words in each tweet.

CB

4. Communicative Behaviour

- Communicative Styles (**CoSty**):
  the fraction of retweets, of replies, and of original tweets over all user's feed.

- Emoji Profile (**EPro**):
  distinguish some user's traits from the emoji her/his used.

Genre: emoji ZWJ sequences 🏊‍♀️ (🏊 Person Swimming, Zero Width Joiner and ♀ Female Sign)

Skin color: emoji ZWJ sequences 🏻 🏼 🏽 🏾 🏿

Religion: religious emojis 🛐 ✝️ ☪️

Nationality: national flag 🇯🇵 🇰🇷 🇩🇪 🇨🇳 🇺🇸 🇫🇷 🇪🇸 🇮🇹 🇷🇺 🇬🇧

Experimental setting

5-fold validation over the train set with the aim of **maximizing** the predictive **accuracy**.

linear Support Vector Machine

The code is available on GitHub:

https://github.com/mirkolai/PAN2021_HaMor

HS MV NE CB

Experimental results

English	
FEATURES	ACCURACY
RSdb, HatEval, FER eMFD, NER,	**73.50%**
RSdb, HatEval, eMFD	71.17%
HatEval, RSdb, NER	70.17%
HatEval, FER	64.17%
ALL	62.72%
BoW	61.50%

Spanish	
FEATURES	ACCURACY
HL, eMFD, BoW	82.83%
HL, NoS, ATA, eMFD, NER, BoW, BoE	**80.98%**
ATA, BoW	79.50%
BoW	77.33%
ALL	77.84%
NoS, ATA, NER, BoE	68.33%

HS MV NE CB

Experimental result

English	
FEATURES	ACCURACY
RSdb, HatEval, FER eMFD, NER,	73.50%

Spanish	
FEATURES	ACCURACY
HL, NoS, ATA, eMFD, NER, BoW, BoE	80.98%

Official results 73% (19th position - over 66 participating teams)

English		
FEATURES	ACCURACY	Rank
RSdb, HatEval, FER eMFD, NER,	62% ↓	43th

Spanish		
FEATURES	ACCURACY	Rank
HL, NoS, ATA, eMFD, NER, BoW, BoE	84% ↑	2nd

HS | MV | NE | CB

Discussion

English		
FEATURES	ACCURACY	Rank
RSdb, HatEval, FER eMFD, NER,	62%	43th

Spanish		
FEATURES	ACCURACY	Rank
HL, NoS, ATA, eMFD, NER, BoW, BoE	84%	2nd

1. Hate Speech Detection

the use of models trained with **HatEval** seems to help in detection Hate Speech Spreaders on Twitter (**HatEval** and **Fermi** for English and **Atalaya** for Spanish)

The lexica of hateful words contribute to reach this level of accuracy (**Racial Slur Database** for English and **Hurtlex** and **No Swearing** for Spanish)

HS | **MV** | **NE** | CB

Discussion

English		
FEATURES	ACCURACY	Rank
RSdb, HatEval, FER eMFD, NER,	62%	43th

Spanish		
FEATURES	ACCURACY	Rank
HL, NoS, ATA, eMFD, NER, BoW, BoE	84%	2nd

2. Moral Values Detection

3. Named Entity Recognition of Hate Speech target

both models benefit of the **External Moral Values dictionary** and of the feature base on **Named Entity Recognition**

HS MV NE **CB**

Discussion

English		
FEATURES	ACCURACY	Rank
RSdb, HatEval, FER eMFD, NER,	62%	43th

Spanish		
FEATURES	ACCURACY	Rank
HL, NoS, ATA, eMFD, NER, BoW, BoE	84%	2nd

4. Communicative Behaviour

the model for English laks of features based on communicative behaviour

Bag of Word (**BoW**)

Bag of Emojy (**BoE**)

Open questions

Did **English** subtask get affected by the laking of features based on **Communicative Behaviour**?

CB + **English = ?**

Does the set of features used for the **Spanish** subtask be proficiently applied to the **English** dataset?

English	
FEATURES	ACCURACY
HL, NoS, ATA, eMFD, NER, BoW, BoE	?

we will verify that when the gold test will be available