

HaMor at the Profiling Hate Speech Spreaders on Twitter

Notebook for PAN at CLEF 2021

Mirko Lai¹, Marco Antonio Stranisci¹, Cristina Bosco¹, Rossana Damiano¹ and Viviana Patti¹

¹*Dipartimento di Informatica, Università degli Studi di Torino, Italy*
The first two authors equally contributed to this work.

Abstract

In this paper we describe the Hate and Morality (HaMor) submission for the *Profiling Hate Speech Spreaders on Twitter* task at PAN 2021. We ranked as the 19th position - over 66 participating teams - according to the averaged accuracy value of 73% reached by our proposed models over the two languages. We obtained the 43th higher accuracy for English (62%) and the 2nd higher accuracy for Spanish (84%). We proposed four types of features for inferring users attitudes just from the text in their messages: HS detection, users morality, named entities, and communicative behaviour. The results of our experiments are promising and will lead to future investigations of these features in a finer grained perspective.

Keywords

Hate Speech, Moral Values, Communicative Behaviour, Named Entities

1. Introduction

The Profiling Hate Speech (HS) Spreaders on Twitter is an Author Profiling task organized at PAN [1, 2, 3]. Teams are invited to develop a model that, given a Twitter feed of 200 messages, determines whether its author spreads hatred contents. The task is multilingual, and covers Spanish and English languages. The training set is composed of 200 users per language, 100 of them annotated as haters by having posted at least one HS in their feeds; the annotation of single tweets is not available, though. All the information about users, mentions, hashtags, and urls are anonymized, making not replicable in this context approaches based on demographic features [4], or community detection [5].

The Hate and Morality (HaMor) team participates to the task with a system that combines HS, and moral values detection [6] in a feed of tweets, in order to infer a general attitude of a user towards people vulnerable to discrimination. Our approach relies on the moral pluralistic

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania


✉ mirko.lai@unito.it (M. Lai); marcoantonio.stranisci (M. A. Stranisci); cristina.bosco@unito.it (C. Bosco); rossana.damiano@unito.it (R. Damiano); viviana.patti@unito.it (V. Patti)

🌐 <http://www.di.unito.it/~lai/> (M. Lai); <http://www.di.unito.it/~stranisci/> (M. A. Stranisci); <http://www.di.unito.it/~bosco/> (C. Bosco); <http://www.di.unito.it/~rossana/> (R. Damiano); <https://www.unito.it/persona/vpatti> (V. Patti)

🆔 0000-0003-1042-0861 (M. Lai); 0000-0001-9337-7250 (M. A. Stranisci); 0000-0002-8857-4484 (C. Bosco); 0000-0001-9866-2843 (R. Damiano); 0000-0001-5991-370X (V. Patti)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

hypothesis (Cfr [7, 8, 9]), according to which moral foundations are many and people prioritize some values than other ones. This can lead to divergent and often conflicting points of view on debated facts, and might also be a factor in HS spreading [10]. More specifically, we considered a group-bound moral judgement as the signal of a potential negative stance against minorities, and used it as a feature to classify HS spreaders together with a HS detection model. The paper is structured as follow: in Section 2 features selection stage is described, and Section 3 is devoted to present the experimental results of our model. In Conclusions (Section 4) some limitations of our approach are discussed.

2. Feature Selection

Four types of features for inferring users attitudes just from the text in their messages have been selected to train our model: HS detection (Section 2.1), users morality (Section 2.2), Named Entities (Section 2.3), Communicative behavior (Section 2.4).

2.1. Hate Speech Detection

Hate Speech Detection (HSD) is the automated task of detecting whether a piece of text contains hate speech. Several shared tasks on HSD have taken place and large annotated corpora are available in different languages. For example, the *HatEval* dataset for hate speech detection against immigrants and women in Spanish and English tweets have been released to be used at the Task 5 of the SemEval-2019 workshop [11]. We decided to use the entire *HatEval* dataset for training three models and we proposed the following features:

- SemEvalSVM (*SESVM*): 1-dimensional feature that counts - for each user - the number of hateful tweets predicted by a linear SVM trained using a text 1-3grams bag-of-words representation.
- Atalaya (*ATA*) [12]: 1-dimensional feature that counts - for each user - the number of hateful tweets predicted by a linear-kernel SVM trained on a text representation composed of bag-of-words, bag-of-characters and tweet embeddings, computed from fastText word vectors. We have taken inspiration from the system proposed by the Atalaya team that obtained the best result at the *HatEval* Spanish sub-task.
- Fermi (*FER*) [13]: a 1-dimensional feature that counts - for each user - the number of hateful tweets predicted by SVM with the RBF kernel trained on tweet embeddings from Universal Sentence Encoder. We based the system on that proposed by the Fermi team that obtained the best result at the *HatEval* English sub-task.

Furthermore, the growing interest on this topic leads the research community (and not only) to develop some lexica of hateful words such as HurtLex [14], NoSwearing¹, and The Racial Slur Database². HurtLex is a lexicon of offensive, aggressive, and hateful words in over 50

¹<https://www.noswearing.com/>

²<http://www.rsd.db.org/full>

languages (including English and Spanish). The words are divided into 17 different categories. Then, NoSwearing is a list of English swear words, bad words, and curse words. The Spanish translation was made by Pamungkas et. al [15]. Finally, the Racial Slur Database is a list of words that could be used against someone - of a specific race, sexuality, gender etc. - divided into more than 150 categories. The list is only available in English, we thus computed the Spanish translation using Babelnet's API [16]. We also take advantage of spaCy³ models *en_core_web_lg*, and *es_core_news_lg* for expanding the three lexica. Indeed, we used the tok2vec embedding representation for including in the three lists the 10 most similar tokens of each word in order to also capture inflection forms and synonyms. We have thus proposed the following features:

- HurtLex (*HL*): a 18-dimensional feature that evaluates the number of hateful words used by each user, the mean of hateful words in each tweet, and the standard deviation. We exploited the following 6 categories: negative stereotypes ethnic slurs, moral and behavioral defects, words related to prostitution, words related to homosexuality, words related to the seven deadly sins of the Christian tradition, felonies and words related to crime and immoral behavior (we exclusively considered the conservative level).
- No Swearing (*NoS*): a 3-dimensional feature that evaluates the number of swearing words used by each user, the mean of swearing words in each tweet, and the standard deviation.
- The Racial Slur Database (*RSdb*): a 27-dimensional feature that evaluates the number of swearing words used by each user, the mean of swearing words in each tweet, and the standard deviation for each of the following 9 categories: Asians, Arabs, Black people, Chinese, Hispanics, Jews, Mexicans, Mixed Races, Muslims.

2.2. Moral Values Detection

According to many scholars, morality is a pluralistic rather than an universal concept. Several configuration of values are possible, and some of them are in conflict, such as autonomy *versus* community [7], or conservation *versus* openness to change [8]. The Moral Foundation Theory (MFT) [9] shares this approach since it distinguishes five dyads leading to people morality: care/harm, fairness/cheating, which relies on individualization, and loyalty/betrayal, authority/subversion and purity/degradation, which are binding foundations. Some of these combinations may correlate with specific political positions, as emerges from experimental results [17]: liberals seem to agree on individualization values, whereas conservatives could be more likely to follow binding dyads.

In building our model, we considered binding moral dyads as a potential feature characterizing a HS spreader. More specifically, we claimed that users who rely on loyalty/betrayal and authority/subversion might be inclined to post hatred contents online. Hence, we referred to two existing resources: the extended Moral Foundations Dictionary (eMFD) [18], and the Moral Foundations Twitter Corpus (MFTC) [6].

The eMFD is a dictionary of terms categorized by a specific moral foundation. We chose all those related to loyalty/betrayal and authority/subversion moral concerns, and translated them in

³<https://spacy.io/>

Spanish with the BabelNet’s API. Finally, we expanded the words list using the same methodology explained in Section 2.1. The result is the following feature:

- extended Moral Foundations Dictionary (*eMFD*): a 12-dimensional feature that comprises the mean, the standard deviation, and the total amount of terms occurring in her/his tweets for the four categories loyalty/betrayal and authority/subversion.

The MFTC is a collection of 35,000 tweets annotated for their moral domains, and organized in 7 subcorpora, each focusing on a specific discourse domain (eg: the Black Lives Matters, and #metoo movements, and the US 2016 presidential elections). Using transfer learning as a label assignment method, we converted the original multi-label annotation schema in a binary-label one: 9,000 texts annotated as loyalty, betrayal, authority or subversion were considered as potentially correlated with HS (*true*), while the other not (*false*). Using the resulting corpora as training set, we thus proposed the following feature.

- Moral Foundations Twitter Corpus (MFTC): a 1-dimensional feature that counts - for each user - the number of hateful tweets predicted by a linear SVM trained using a text 1-3grams bag-of-words representation.

2.3. Named Entity Recognition of HS target

In a message, the mention of a person belonging to a group vulnerable to discrimination might be seen as a signal of hatred contents, since the clear presence of a target in this kind of expressions allows discriminating between what is HS and what is not. Thereby, we implemented a feature aimed at detecting the presence of a potential HS target within a tweet.

We first collected all the entities of type PERSON in the whole training set detected by the transition-based named entity recognition component of spaCy. Then, we searched the retrieved entities on Wikipedia through the Opensearch API⁴. The example below shows the Wikipedia pages returned by the Opensearch API when the entity *Kamala* is requested.

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

However, this operation is revealed to be not accurate. In fact, it does not return a unique result for each entity detected by spaCy, but a set of 10 potential candidates. Therefore, we decided to create two lists - one for each language - of HS targets including only persons that belong to categories that could be subject to discrimination.

With the aim of detecting the relevant categories, we scraped the *category box* from the Wikipedia pages of all entities of type PEOPLE detected by spaCy (3,996 English, and 5,089 Spanish). The result is a list of Wikipedia’s categories per language, which needed to be filtered to avoid not relevant results.

⁴<https://www.mediawiki.org/wiki/API:Opensearch>

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](#)

Figure 1: A selection of categories for Kamala Harris on Wikipedia’s category box

The Figure 1 shows a partial selection of *Kamala Harris* category box, which contains several references to unnecessary information, such as ‘1964 births’, or ‘Writers from Oakland, California’, but also usefully ones, such as ‘*African-American* candidates for President of the United States’ or ‘*Women* vice presidents’.

After a manual analysis of the two lists, we thus narrowed them by a regex filtering, in order to obtain only a set of relevant categories: 279 for English, and 415 for Spanish. Finally, we collected all the individuals who are their members. As final result, we obtained two gazetteers of potential HS targets (7, 5890 entities for English, and 31, 235 for Spanish) in the following format.

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

We thus proposed a feature that counts the mentions towards persons belonging to a group vulnerable to discrimination.

- Named Entity Recognition of HS target (NER): a 5-dimensional feature expressing the total number of potential HS targets mentioned in her/his tweets, the mean, the standard deviation, and the ratio between the number of HS target, and all the HS targets mentioned by the user.

2.4. Communicative Behavior

Under the label ‘Communicative Behavior’ a set of features related to the structure of the tweet and to the user’s style has been grouped. The total number, the mean, and the standard deviation have been computed for each feature over all users feeds.

- Uppercase Words (UpW): this feature refers to the amount of words starting with a capital letter and the number of words containing at least two uppercase characters.
- Punctuation Marks (PM): a 6-dimensional feature that includes 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.
- Communicative Styles (CoSty): a 3-dimensional feature that computes the fraction of retweets, of replies, and of original tweets over all user’s feed.
- Emoji Profile (EPro): this feature tries to distinguish some user’s traits from the emoji her/his used. We implemented a one-hot encoding representation of the modifiers used in the emoji ZWJ sequences (e.g. *man: medium skin tone*, beard) that includes the 5 different skin tone modifiers and the gender modifiers, in addition to the religious emojis (e.g. Christian Cross) and the national flags.

We finally employed bag-of-words models as feature:

- Bag of Words (BoW): binary 1-3grams of all user’s tweets.
- Bag of Emojis (BoE): binary 1-2grams of all user’s tweets only including emojis.

3. Experiments and Results

The organizers provided a train dataset of 400 Twitter’s feeds - 200 written in English and 200 in Spanish - binary labelled with HS Spreader. The distribution is perfectly balanced among the true and false labels. In order to assess the performance of the participating systems, a test set of 200 unlabelled Twitter’s feeds - 100 for each language - was provided. The accuracy was used for evaluating the performance in terms of HS Spreader identification for each language. The final ranking will be the average of the accuracy values per language.

3.1. HaMor experiments

In our experiments, we addressed HS Spreader identification employing a 5-fold validation over the train set with the aim of maximizing the predictive accuracy. The code is available on GitHub for further exploration and for allowing reproducibility of our experiments⁵. We carried out several experiments by combining all the features introduced in Section 2 for training a linear Support Vector Machine (SVM). Table 1 shows some of the performances we obtained in this experimental setting for each languages. The values for the accuracy are the average over three 5-fold validations using different random shuffles of the train set. The runs we submitted have been highlighted in bold print.

The first row contains the higher result that we obtained in the experimental setting for the two languages. Despite this, we have chosen to submit a Spanish run that reached a lower result. We experimented in other shared tasks that the results obtained in the experimental setting is often much lower than the official ones [19]. It was precisely for this reason that we chose features - that reached good results, though not the best - based on external sources such as other annotated corpora (FER, ATA, HatEval), lexica (HL, eMFD, NoS), and semi-structured contents (NER). Indeed, if, on one hand, the submitted model for English does not reached the

⁵https://github.com/mirkolai/PAN2021_HaMor

Table 1
Experimental Results on the Training set

ENGLISH		SPANISH	
FEATURES	ACCURACY	FEATURES	ACCURACY
<i>NER, eMFD, RSdb, HatEval, FER</i>	73.50	<i>BoW, eMFD, HL</i>	82.83
<i>eMFD, RSdb, HatEval</i>	71.17	<i>BoW, BoE, NER, eMFD, HL, NoS, ATA</i>	80.98
<i>NER, HatEval, RSdb</i>	70.17	<i>BoW, ATA</i>	79.50
<i>HatEval, FER</i>	64.17	<i>BoW</i>	77.33
<i>ALL</i>	62.72	<i>ALL</i>	77.84
<i>BoW</i>	61.50	<i>NER, BoE, NoS, ATA</i>	68.33

higher level of accuracy reached in the experimental setting, on the other hand, the submitted model for Spanish reached a better result than experimental ones. The omission of BoW and BoE features adversely affects the experimental runs for Spanish, but not for English. However, our choice of not include the features BoW and BoE in the submitted run for English involves the creation of a little feature-space representation of each instance (vectors length of 52 variables). This representation could indeed include too few variables for effectively detecting a so complex phenomena such as HS. Then, we did not investigated cross-cultural differences among English and Spanish speaking countries although it is well known that historical and political factors are relevant on the orientation of moral values [20]. Maybe also for that reason, the features based on the moral values we investigated performs better in Spanish then in English tweets. However, the introduction of the external-sources-based features positively affects the runs for both languages.

3.2. Official results

Our models obtained 84% (2nd higher result) and 62% (43th higher result) in terms of accuracy on HS Spreader identification respectively for Spanish and English. The final score, used in determining the the final ranking, is the averaged accuracy values per language which corresponds to 73% (19th position - over 66 participating teams).

4. Conclusions

In this paper we presented an overview of the HAMOR submission for the *Profiling Hate Speech Spreaders on Twitter* task at PAN-2021. We participated by submitting one run for both tweets written in English and Spanish. Our approach, chiefly based on external resources such as other annotated corpora, lexica, and semi-structured content, proved to be highly successful concerning the task of HS Spreader identification in both languages. The results show that the use of external resources preserves stable values of accuracy between the experimental setting and the prevision of the test set on Spanish sub-task. The proposed lexica gave a considerable contribution for obtaining these results and the use of named entity recognition for detection potential target of HS looks promising. In the future, we plan to employ the features discarded from the submitted run for a prediction on the test set. We also aim to explore a finer grained approach to MFT detection features, considering different combination of moral values, and analyzing how moral attitudes

may vary across different countries. Finally, the Named Entity Recognition feature needs to be improved through testing different NER tools, and referring to other semantic resources, in addition to Wikipedia.

References

- [1] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.
- [2] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
- [3] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.
- [4] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, 2016, pp. 88—93.
- [5] P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, Author profiling for abuse detection, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1088–1098.
- [6] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaladar, 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.
- [7] R. A. Shweder, N. C. Much, M. Mahapatra, L. Park, The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering., *Morality and Health* (1997) 119–169.
- [8] S. H. Schwartz, An overview of the Schwartz theory of basic values, *Online readings in Psychology and Culture* 2 (2012) 2307–0919.
- [9] 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.
- [10] J. Hoover, M. Atari, A. M. Davani, B. Kennedy, G. Portillo-Wightman, L. Yeh, D. Kogon, M. Dehghani, Bound in hatred: The role of group-based morality in acts of hate, *PsyArXiv* (2019).
- [11] 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 Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.

- [12] 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.
- [13] 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.
- [14] E. Bassignana, V. Basile, V. Patti, Hurltlex: 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.
- [15] E. W. Pamungkas, A. T. Cignarella, V. Basile, V. Patti, et al., 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets, in: 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018, volume 2150, CEUR-WS, 2018, pp. 234–241.
- [16] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial intelligence* 193 (2012) 217–250.
- [17] J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations., *Journal of personality and social psychology* 96 (2009) 1029.
- [18] 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.
- [19] M. Lai, A. T. Cignarella, D. I. Hernández Farías, iTACOS at IberEval2017: Detecting stance in Catalan and Spanish tweets, in: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), volume 1881 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 185–192.
- [20] L. Myyry, K. Helkama, M. Silfver-Kuhlampi, K. Petkova, J. P. Valentim, K. Liik, Explorations in reported moral behaviors, values, and moral emotions in four countries, *Frontiers in Psychology* 12 (2021) 1468.