Analysing Moral Beliefs for Detecting Hate Speech Spreaders on Twitter

Mirko Lai(✉), Marco Antonio Stranisci, Cristina Bosco, Rossana Damiano, and Viviana Patti

Dipartimento di Informatica, Università degli Studi di Torino, Turin, Italy
{mirko.lai,marco.stranisci,cristina.bosco,rossana.damiano,
viviana.patti}@unito.it

Abstract. The Hate and Morality (HaMor) submission for the *Profiling Hate Speech Spreaders on Twitter* task at PAN 2021 ranked as the 19th position - over 67 participating teams - according to the averaged accuracy value of 73% over the two languages - English (62%) and Spanish (84%). The method 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. In this paper, since the test set is now available, we were able to analyse false negative and false positive prediction with the aim of shed more light on the hate speech spreading phenomena. Furthermore, we fine-tuned the features based on users morality and named entities showing that semantic resources could help in facing Hate Speech Spreaders detection on Twitter.

Keywords: Hate Speech · Moral Foundation Theory · Twitter

1 Introduction

The Profiling Hate Speech (HS) Spreaders on Twitter is an Author Profiling task [17] organised at PAN [4]. 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 anonymised, making not replicable in this context approaches based on demographic features [22], or community detection [3, 13].

Our team participated to the task with a system called *The Hate and Morality* (HaMor). The name of the model refers to the combined use of HS and moral values detection [7] for analysing 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

M. Lai and M. A. Stranisci—Contributed equally to this work.

A. Barrón-Cedeño et al. (Eds.): CLEF 2022, LNCS 13390, pp. 149–161, 2022.
https://doi.org/10.1007/978-3-031-13643-6_12

pluralistic hypothesis (Cfr [6,18,19]), according to which moral foundations are many and people more prioritise 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 [8]. 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: Sect. 2 brings again the attention on the description of the features used in the task, and Sect. 3 is devoted to an error analysis focusing on a better understanding of false positive cases. Section 4 proposes a qualitative analysis of the proposed features. Then, Sect. 4.3 describes the improvements made to our system for better predicting HS Spreaders on Twitter. In Conclusions (Sect. 5) the contribution of our approach on this phenomena 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 (Sect. 2.1), users morality (Sect. 2.2), Named Entities (Sect. 2.3), Communicative behaviour (Sect. 2.4). We employed a manual ensemble-based feature selection method combining multiple feature subsets for selecting the optimal subset of features that improves classification accuracy for each language.

2.1 Hate Speech Detection

HS detection is the automated task of detecting whether a piece of text contains HS. Several shared tasks on HD detection 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* has been released to be used at the Task 5 of the SemEval-2019 workshop [1]. 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 1–3 g bag-of-words representation.
- Atalaya (*ATA*) [16]: 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 were inspired from the Atalaya team's system that achieved the best scores in the *HatEval* Spanish sub-task.
- Fermi (*FER*) [10]: 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 were inspired by the Fermi team's system 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 [2], NoSwearing[1], and The Racial Slur Database[2]. HurtLex is a lexicon of offensive, aggressive, and hateful words in over 50 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, sex, gender etc. - divided into more then 150 categories. The list is only available in English, we thus computed the Spanish translation using Babelnet's API [14]. We also take advantage of spaCy[3] 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. We can thus propose 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 behavioural 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 behaviour (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, moral beliefs are not universal, but reside on a plurality of "irreducible basic elements" [21]. Several configuration of values are possible, and some of them are in conflict, such as autonomy *versus* community [19], or conservation *versus* openness to change [18]. The Moral Foundation Theory (MFT) [6] shares this approach since it distinguishes five dyads leading to people morality: care/harm, fairness/cheating, which relies on individualisation, 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 [5]: liberals seem to agree on individualisation values, whereas conservatives could be more likely to follow binding dyads.

In building our model, we considered binding moral dyads as a potential feature characterising 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.

[1] https://www.noswearing.com/.

[2] http://www.rsdb.org/full.

[3] https://spacy.io/.

Hence, we referred to two existing resources: the extended Moral Foundations Dictionary (eMFD) [9], and the Moral Foundations Twitter Corpus (MFTC) [7].
The eMFD is a dictionary of 2,965 terms categorised by a specific moral foundation. We chose all those related to loyalty/betrayal and authority/subversion moral concerns, and translated them in Spanish scripting babbel.com and wordreferences.com (the translated dictionary amounts to 4,622 words). Finally, we expanded the words list using the same methodology explained in Sect. 2.1. The result is the following feature: for each user we computed the mean, the standard deviation, and the total amount of terms occurring in her/his tweet.

- extended Moral Foundations Dictionary (*eMFD*): a 12-dimensional feature that includes 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 organised in 7 subcorpora, each focusing on a specific discourse domain (e.g.: 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 1–3 g 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 API[4]. The example below shows the Wikipedia pages returned by the Opensearch API when the entity *Kamala* is requested.

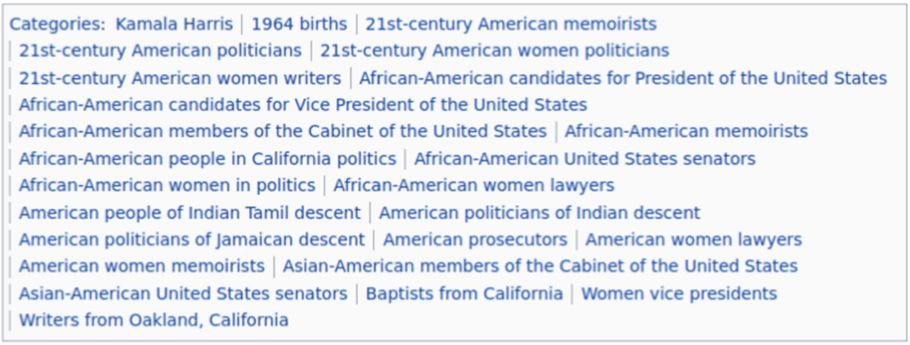
```
['Kamala','Kamala Harris','Kamal Haasan',
'Kamala (wrestler)','Kamala Khan','Kamala Surayya',
'Kamala Harris 2020 presidential campaign',
'Kamaladevi Chattopadhyay','Kamala Mills fire',
'Kamalani Dung']
```

[4] https://www.mediawiki.org/wiki/API:Opensearch.

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.

The Fig. 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 us '*African-American* candidates for President of the United States' or '*Women* vice presidents'.



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

Fig. 1. A selection of categories for Kamala Harris on Wikipedia's category box

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 Behaviour

Under the label 'Communicative behaviour' 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 1–3 g of all user's tweets.
– Bag of Emojis (BoE): binary 1–2 g of all user's tweets only including emojis.

3 Error Analysis

The organisers provided a dataset for training participant systems including 400 Twitter's feeds - 200 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 also provided.

The current availability of the correct labels for the test set allows us to perform an error analysis that we focus on better understanding the false positive cases. The test set is balanced for both languages (50% of the users are hate speech spreaders).

Table 1 shows confusion matrix of our submission for both languages. For each of the languages, the entry in row 0 and column 1 indicates the amount of *false positives*, i.e. samples that our system erroneously predicted as HS spreader (1) while they weren't. The entry in row 1 and column 0 indicates the amount of *false negatives*.

For both languages, the number of false positives is similar to the amount of false negatives, while in Spanish fewer errors in the prediction of HS spreaders can be observed with respect to English.

We aim to perform a manual error analysis mostly evaluating the tweets of the users that are not HS spreaders, but that have been predicted as such by our model. Unfortunately, also observing the correct labels provided by the organizers, we cannot

Table 1. Confusion Matrix

		EN			ES		
		Predicted			***Predicted***		
		0	*1*		*0*	*1*	
Actual	*0*	33	**17**	50	41	**9**	50
	1	**21**	29	50	**7**	43	50
		54	46		48	52	

check whether a single tweet is HS or not, hence labels only indicate whether the user that generated the feed (where the tweet is included) is a hater or not. Since then a user feed is composed by several tweets, we decided to filter them by automatically predicting whether each single tweet is hateful or not using one of the models proposed in Sect. 2.1: SESVM, ATA, and FER.

Figure 2 shows the number of users *y* having at least *x* tweets that have been predicted as hateful in their feeds by our models.

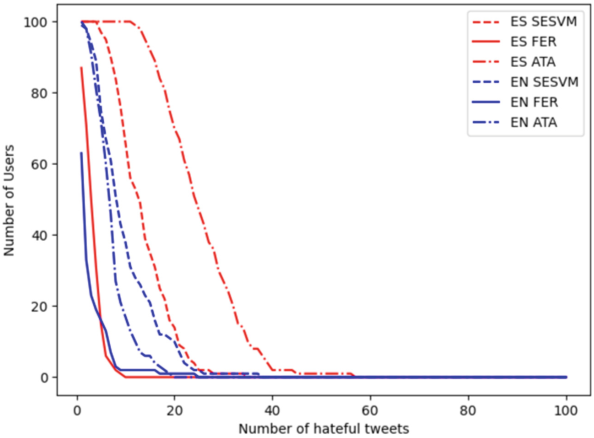


Fig. 2. The number of users *y* having at least *x* hateful tweets

FER is the model that shows a more conservative trend: it predicts not more than one tweet as hateful in 62 English and 84 Spanish users' feeds. Furthermore, it does not predict more than 10 hateful tweets in any Spanish users' feeds and it follows a similar trend on both languages. On the contrary, ATA and SESVM are more inclusive predicting at least 1 hateful tweets in all users' feeds. Furthermore, ATA seems to be more conservative on English than on Spanish tweets. For such reasons, we decided to use FER for automatically predicting HS in individual tweets.

The results provided by FER allows us to better understand the motivation of the erroneous classification of some user as HS spreader, that wasn't according to the correct labels provided by the organisers for the test set. For what concerns English, we find several tweets in hate-less speech spreaders feeds containing profanities and slurs. As an illustrative example, a feed of a black woman that wrote:

> My nigga just came home with a Lush 2[5]. Goodnight bitches 😂

Although the author of the tweet uses emojis that include skin tones and the female sign in her feed, these signs do not help the model to understand that she is a black woman that uses the words *nigga* (racism) and *bitches* (misogyny) in a funny way for communicating with her followers.

Also in some Spanish case, although the false positive entries are very few, we found several profanities and bad words in false positive hate-less speech spreaders feeds. Following an illustrative example of three tweets extracted from the same false positive user's feed:

> "#USER# Para toda la mierda femiorca[6], que os jodan hdps" (*For all the femiorca shit, fuck you son of a beach*)

> Pinta negro para cualquier persona a día de hoy. Esto es vivir en un imposible. #URL# (*Paint black for anyone today. This is living in an impossible. #URL#*)

> Y los que no son junden son masones. Que asco de UE. #URL# (*And those who are not Jew are Masons. What a mess of the EU. #URL#*)

Although the author of the tweet has not been considered a HS spreader by the organisers of the task, these three tweets express very strong and questionable opinions against feminist movements, black people rights, and Jews. For our model it is therefore difficult to not predict this user as a HS spreader.

4 Features Analysis and Improvement

Experimental results showed a significant delta between the two languages, despite both relied on a similar set of features. Hence, in this section we provide a deeper analysis of features adopted in our proposal, with a specific focus on MFT Values, and Named Entities.

4.1 MFT Values

In our experimental setting, we selected only two moral dyads from the MFT. This choice relied on psychological studies claiming for a correlation between the political stance of a person and certain moral configurations. However, such assumption is derived from psychological surveys rather than from NLP experiments. Thereby, we analysed how MFT dimensions correlate with HS spreading. We used the eMFT dictionary [9], to count all the occurrences of words expressing MFT values for each user. Then, we computed the Spearman's correlation between each value and HS spreaders

[5] Lush 2 is a Sex Toys.

[6] Femiorca is a feminist community.

labels in order to observe which were more significant for the task. As it can be observed in Table 2, the role of MFT values may be more relevant for Spanish, since the average ρ for this language is 0.26 while the average ρ for English is 0.09. In both languages there is always one element in each dyad that better correlates with HS. For instance, Harm obtains a higher Spearman's ρ score than Care. This may suggest the existence of a set of different moral frames adopted by users (Cfr [11]) that shape their communicative behaviour. A closer look into the dyads shows some interesting trends about the correlation between moral values and HS. Harm and Subversion are predominant in their respective dyads for English and Spanish, suggesting a moral configuration in which binding and individualisation values interact in determining users stance. On the opposite, the Purity/Degradation dyad behave differently between languages. English HS spreaders seem to focus on the violation of the dyad (Degradation), while Spanish users do the opposite. Finally, none of the Fairness/Cheating and Loyalty/Betrayal values significantly correlates with HS in English. Such distribution seems to confirm that moral stance is topic-sensitive, as demonstrated by [7]. Further investigation in existing corpora may shed more light on this phenomenon.

Table 2. The Spearman's correlation of each MFT values with HS Spreader in the dataset.

Moral Value	**Spearman's ρ (en)**	**Spearman's ρ (es)**
Loyalty	0.003	0.276
Betrayal	0.069	0.134
Purity	0.027	0.406
Degradation	0.181	0.329
Care	-0.035	0.144
Harm	0.137	0.404
Fairness	0.075	0.337
Cheating	0.015	0.038
Authority	0.012	0.174
Subversion	0.143	0.359

We then proposed a feature that includes the full spectrum of moral values: eMFD+.

4.2 Named Entities

In our original submission, the creation of gazetteers with named entities who are potentially target of HS was based on a manual selection of Wikipedia categories containing some target words related to vulnerable groups (e.g.: American women non-fiction writers). This led to sparse representations of this feature, since we obtained $11,480$ categories of people for Spanish, and $36,366$ for English and most of them were not mentioned by users in their tweets. We decided to remove all categories of named entities that were mentioned by less than 20 users in the data set, dramatically reducing the number of categories to 204 for Spanish and 225 for English. Finally, we computed the Spearman's correlation between the occurrences of each category and HS spreaders labels. Table 3 shows the 5 categories which best correlate with HS. As it can be observed, women are a shared target across languages, while religious minorities are a significant target for English and LGBT for Spanish. As for MFT values, it seems that the distribution and relevance of vulnerable categories for HS detection is strongly influenced by current events. For instance, several mentions of Kamala Harris appear to be correlated with the 2020 US elections.

Table 3. The Spearman's correlation of each category of people vulnerable to HS and HS Spreader in the dataset.

Category of people (en)	**Spearman's ρ**	**Category of people (es)**	**Spearman's ρ**
American_women_podcasters	0.200	Feministas_de_Madrid	0.440
American_women_rock_singers	0.189	Mujeres_guerreras_ficticias	0.267
American_women_non-fiction_writers	$0, 175$	Mujeres	0.220
Kenyan_Muslims	0.171	Artistas_LGBT_de_España	0.214
American_women_memoirists	0.165	Mujeres_del_siglo_XX	$0, 206$

We propose an enhanced version of the NER feature (NER+) that exclusively takes in consideration the entities belonging to this filtered set of categories.

4.3 Fine-Tuning

Our official submission obtained 84% and 62% in terms of accuracy on HS Spreader identification respectively for Spanish and English. The final score, used in determining the final ranking, is the averaged accuracy values per language which corresponds to 73% [17]. Here, we verify the contribution of the fine-tuned featured described in Sect. 4.

English. Our submission for the English subtask employed the features NER, eMFD, RSdb, HatEval, and FER. The dimensional space representation of each user's feed was relatively simple and the obtained results was 12% points below the highest one (the UO-UPV [12] team obtained 74%).

Thereby, we tried to increase the complexity representation adding the Communicative Behaviour feature BoW to this configuration. The model achieves 65% in term of accuracy, still very much below the state of the art. We then employed the features NER+, eMFD+, and replaced ATA with FER which has been shown to be more skewed on precision in detecting HS (see Sect. 3). The obtained accuracy increased of other 2%points. Table 4 shows the contribution of each fine-tuned features. Replacing FER with ATA does not affect the result, as well as the enhanced NER feature seems to not improving the prediction. However, the effectiveness of feature based on the full spectrum of moral values (eMFD+) is showed.

Table 4. Evaluation of the contribute of enhanced on English subtask

Feature Set	**Accuracy**
NER, eMFD, RSdb, HatEval, and FER	0.67
replacing NER with NER+	*0.67*
replacing eMFD+ with eMFD	0.63
replacing FER with ATA	0.67

Spanish. For Spanish submission, we employed two Communicative Behaviour features (BoW and BoE), NER, eMFD, HL, NoS, ATA. We then applied the enhanced version of eMFD and NER and we also tried to replace ATA with FER in order to test a more conservative feature. The obtained accuracy increased of 1%point achieving the highest result obtained by the team SiinoDiNuovo [20]. Table 5 shows the contribution of each fine-tuned features:

Table 5. Evaluation of the contribute of enhanced on Spanish subtask

Feature Set	**Accuracy**
BoW, BoE, NER, eMFD, HL, NoS, ATA	0.85
replacing NER with NER+	*0.85*
replacing eMFD+ with eMFD	0.78
replacing FER with ATA	0.83

Also in this case, the effectiveness of feature based on the full spectrum of moral values (eMFD+) is showed. Then, a conservative feature based on HS detection such as FER better affected the result. Finally, we could employed NER+ without making any significant changes.

Cross-Language. We finally have given some thought to how the decision to propose different features set for each language had been a good choice. We therefore trained the English model with the features set used for Spanish employing the enhanced version of NER and eMFD. The performance increased further to 71% accuracy for the English subtask. It would have meant the achievement of 78% average accuracy (85% and 70% respectively for Spanish and English) over the two languages (in other words, 2th position in the official ranking with a detachment of only 1% points from the 1st position).

Therefore, the choice to use different set of features for the two languages was inauspicious. However, the effectiveness of features based on lexica (HL, NoS), morality values (eMFD), and Named Entity Recognition (NER) in a multilingual perspective is therefore confirmed and leaves opportunities for further future exploration open.

5 Conclusions

In this paper we presented a detailed analysis of the HAMOR submission for the *Profiling Hate Speech Spreaders on Twitter* task at PAN-2021. 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, as our system ranked as the 19th position among 67 participating teams. 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 deeper explored the features base on named entity recognition and proposed a finer grained approach for employing MFT features, considering different combination of moral values, and analyzing how moral attitudes may vary across different countries. Finally, we propose a cross lingual set of features that improve the result obtained by our model in term of accuracy. All the code used for on this work is available on GitHub for further exploration and for allowing reproducibility of our experiments[7].

References

1. Basile, V., et al.: 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, pp. 54–63. Association for Computational Linguistics (2019)
2. Bassignana, E., Basile, V., Patti, V.: Hurtlex: a multilingual lexicon of words to hurt. In: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, vol. 2253, pp. 1–6. CEUR-WS (2018)
3. Cignarella, A.T., Lai, M., Bosco, C., Patti, V., Rosso, P.: Sardistance@evalita2020: overview of the task on stance detection in Italian tweets. In: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), CEUR Workshop Proceedings, vol. 2765. CEUR-WS.org, Aachen (2020)
4. Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.): Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021, CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
5. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. J. Pers. Soc. Psychol. **96**(5), 1029 (2009)
6. Haidt, J., Joseph, C., et al.: The moral mind: how five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. Innate Mind **3**, 367–391 (2007)
7. Hoover, J., et al.: Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiment. Social Psychol. Pers. Sci. **11**(8), 1057–1071 (2020)
8. Hoover, J., et al.: Bound in hatred: the role of group-based morality in acts of hate. PsyArXiv (2019)

[7] https://github.com/mirkolai/PAN2021_HaMor.

9. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended Moral Foundations Dictionary (eMFD): development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behav. Res. Methods **53**(1), 232–246 (2021)
10. Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., Varma, V.: 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, pp. 70–74. Association for Computational Linguistics (2019)
11. Kwak, H., An, J., Jing, E., Ahn, Y.Y.: Frameaxis: characterizing microframe bias and intensity with word embedding. PeerJ Comput. Sci. **7**, e644 (2021)
12. Labadie, R., Castro-Castro, D., Ortega Bueno, R.: Deep modeling of latent representations for twitter profiles on hate speech spreaders identification: notebook for PAN at CLEF 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021, CEUR Workshop Proceedings, vol. 2936, pp. 2035–2046. CEUR-WS.org (2021)
13. Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E.: Author profiling for abuse detection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1088–1098 (2018)
14. Navigli, R., Ponzetto, S.P.: Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
15. Pamungkas, E.W., Cignarella, A.T., Basile, V., Patti, V., 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, vol. 2150, pp. 234–241. CEUR-WS (2018)
16. Pérez, J.M., Luque, F.M.: Atalaya at SemEval 2019 task 5: robust embeddings for tweet classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 64–69. Association for Computational Linguistics (2019)
17. Rangel, F., De la Peña Sarracén, G.L., Chulvi, B., Fersini, E., Rosso, P.: Profiling hate speech spreaders on twitter task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021, CEUR Workshop Proceedings, vol. 2936, pp. 1772–1789. CEUR-WS.org (2021)
18. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online Read. Psychol. Cult. **2**(1), 2307–0919 (2012)
19. Shweder, R.A., Much, N.C., Mahapatra, M., Park, L.: The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. Morality Health, 119–169 (1997)
20. Siino, M., Di Nuovo, E., Tinnirello, I., La Cascia, M.: Detection of hate speech spreaders using convolutional neural networks. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021, CEUR Workshop Proceedings, vol. 2936, pp. 2126–2136. CEUR-WS.org (2021)
21. Stranisci, M., De Leonardis, M., Bosco, C., Viviana, P.: The expression of moral values in the twitter debate: a corpus of conversations. IJCoL - Special Issue: Comput. Dial. Model. Role Pragmatics Common Ground Interact. **7**(1,2), 113–132 (2021)
22. Waseem, Z., Hovy, D.: 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)