Stance or insults?

Simona Frenda
Dipartimento di Informatica
Università degli Studi di Torino, Turin, Italy
PRHLT Research Center
Universitat Politècnica de València, València, Spain
frenda@di.unito.it

Noriko Kando
National Institute of Informatics
Tokyo, Japan
kando@nii.ac.jp

Viviana Patti
Dipartimento di Informatica
Università degli Studi di Torino, Turin, Italy
patti@di.unito.it

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València, València, Spain
prosso@dsic.upv.es

ABSTRACT

Important issues, such as abortion governmental laws, are discussed everyday online involving different opinions that could be favorable or not. Often the debates change tone and become more aggressive undermining the discussion. In this paper, we analyze the relation between abusive language and the stances of disapproval toward some controversial issues that involve specific groups of people (such as women), which are commonly also targets of hate speech. We analyzed the tweets about the feminist movement and the legalization of abortion events released by the organizers of Stance Detection shared task at SemEval 2016. An interesting finding is the usefulness of semantic and lexical features related to misogynistic and sexist speech which improve considerably the sensitivity of the system of stance classification toward the feminist movement. About the abortion issue, we found that the majority of the expressions relevant for the classification are negative and aggressive. The improvements in terms of precision, recall and f-score are confirmed by the analysis of the correct predicted unfavorable tweets, which are featured by expressions of hatred against women. The promising results obtained in this initial study demonstrate indeed that disapproval is often expressed using abusive language. It suggests that the monitoring of hate speech and abusive language during the stance detection process could be exploited to improve the quality of the debates in social media.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Retrieval tasks and goals**; **Information extraction**; • **Computing methodologies → Natural language processing**; **Information extraction**;

KEYWORDS

Stance Detection, Abusive Language, Sexist and Misogynistic Speech, Social Media

ACM Reference Format:
Simona Frenda, Noriko Kando, Viviana Patti, and Paolo Rosso. 2019. Stance or insults?. In *The Ninth International Workshop on Evaluating Information Access (EVIA 2019), June 10, 2019 at National Institute of Informatics, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Various important issues are discussed everyday online by several users and considering the big amount of shared data online, the possibility of analyzing them to access some specific information became an important task for companies as well as for political organizations. In the Big Data era, for instance, political institutions by means of correct interpretation of data could understand users' opinions about individuals (especially candidates in election campaign periods) or about some controversial issues, in order to provide regulations or measures that could be more favorably accepted by public opinion.

In this perspective, stance detection analyses are increased in the recent years exploring public opinion about different targets on various genres of text. Automatic stance detection aims to determine whether the author of the text is in favor or against toward a given target. However, especially for important social issues, such as laws to permit abortion, the tone of the discussion often become aggressive and offensive:

> (1) @Fungirl3part2 repent wen u commit a grave act like murder of a baby did u #abort ur baby?yes? then YOU repent! #hell s 4 eternity #abortion

> (2) One day I'm gonna set an abortion clinic on fire. Anyone wanna join? #prolife

> (3) Now, I understand your a feminist and think that's adorable, but this grow up time and I'm the man here so run along.

NTCIR '19, June 10-13, 2019 , Tokyo, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

> (4) MARRIAGE for a man is MURDERAGE, That's right MURDER'RAGE! Women have ruined the trust of men, and destabilized their own future. #feminism[1]

The tweets (1) and (2) do not express only a disapproval toward the legalization of the abortion, but hurt individuals and incite hatred and violence. As well as the tweets (3) and (4) are clear examples of expression of sexist and misogynistic opinions.

Therefore, the possibility to capture hate speech especially in opinions that disagree the targeted issue could help the social platforms to improve the quality of debates and avoid the spread of hateful contents online, that amplifies social misbehaviors. With this purpose, in this paper we propose a novel approach to detect stance toward controversial social issues investigating the effectiveness of features able to capture abusive language and aggressive attacks.

In particular, we focused on the topics of the feminist movements and the legalization of abortion, that are ever active issues of discussion. We think that for the detection of the stance toward these political and social controversial issues, dedicated approaches could improve the performance of the classification.

Considering these two topics, we took into account the presence of offensive and hateful expressions against women especially in unfavorable opinions. Therefore, we implemented a computational model to detect stance on Twitter using features able to capture the style and relevant expressions, and lexical and semantic information concerning specifically misogynistic and sexist speech. We approached this task as a classification problem, predicting the higher probability of a tweet to belong to the "against", "favor" or "neutral" class.

In order to evaluate the contribution of these features, we analyzed precision, recall and f-score measures, typically used to evaluate information retrieval tasks. By means of these measures, we can affirm the validity of our approach. As benchmark corpora, we used the training and test sets about the feminist movement and the legalization of abortion released by the organizers of Stance Detection shared task in SemEval 2016[2] [10]. Comparing the performance of our model with the participating systems, our models reach the highest result in the classification of the stance toward legalization of abortion, and overcome the challenging baselines in the detection of stance toward the feminist movement.

Therefore, the main contributions of our work are:

i. introducing a novel way to approach stance detection, showing that features aiming to capture abusive expressions improve the performances of stance detection systems;

ii. investigating the important role of lexical features and the contribution of semantic information for stance detection toward sensitive and controversial issues.

The rest of the paper is structured as follows. The next section outlines the related work. Section 3 describes the corpora and their analyses. Section 4 and 5 describe the used approach focusing on feature engineering and experiments. Section 6 explains the evaluation metrics and the obtained results. Finally, Section 7 and 8 discuss the obtained results and draw some conclusions, proposing a plan for future works.

2 RELATED WORK

In the last years, the Sentiment Analysis field in Natural Language Processing studies is branched out in different specific fields of research, investigating various aspects of political and social communication especially online. Moreover, the growing interest in information access in user-generated contents is supported by national and international campaigns that allow to share data as benchmarks to compare different approaches, such as SemEval.

Regarding Stance Detection task, the authors of [10] proposed for the first time, in SemEval edition of 2016, a shared task asking participant systems to classify whether the tweeter is in favor or against the given target, or whether neither inference is likely. The organizers provided stance data from English Twitter for 6 targets (atheism, climate, feminism, abortion, Hillary's and Trump's campaigns). Proposing a real world challenge, the target could be or not be referred to in the tweet, and sometimes, the target of the opinions is not the pre-chosen target but the competitive entity.

For the purpose of this work, we extracted from this dataset the tweets related to two of the proposed targets: feminist movement and legalization of the abortion.

On this task, various studies are proposed investigating different aspects involved in the stance of the user toward a target. Some researchers analyzed the role of social relations on social platforms [9, 13]; others focused more on sentiment and emotional analyses including word and character n-grams [11], or on structural features (mentions and hashtags) and context-based information [8], exploring supervised [7] and unsupervised approaches [14].

Moreover, given the purpose of this study, some works about aggressiveness and offenses detection online provide useful inspiration. Considering the topics of our investigation, we rely on some previous investigations about misogyny and sexism detection.

To our knowledge, the work in [1] is the first study to face the problem of misogyny identification. The authors created a corpus of English tweets that was used as training and test sets of Automatic Misogyny Identification (AMI) share task [4, 5] at IberEval 2018[3] and EvalIta 2018[4]. The authors compared the performance of different supervised approaches using word embeddings, stylistic and syntactic features.

With respect to sexism, the authors of [15] created a corpus of English tweets NAACL_SRW_2016_tweets[5] annotated with "sexist", "racist" and "none" labels. On this corpus they explored the performance of unigrams, bigrams, trigrams, and fourgrams in a logistic regression based model.

Although this work is inspired by these previous researches, its main scope is to analyze the presence of hate speech especially in opinions that disapprove a specific issue, and understand the advantages of its identification in the stance detection process.

[1] These tweets are extracted from the dataset released by the organizers [10] of Stance Detection shared task in SemEval 2016.
[2] http://alt.qcri.org/semeval2016/task6/
[3] https://amiibereval2018.wordpress.com/
[4] https://amiEvalIta2018.wordpress.com/
[5] The NAACL_SRW_2016_tweets corpus is available online: https://github.com/ZeerakW/hatespeech

3 DATASETS

The datasets used in this work contains the tweets targeting the feminist movement and the legalization of the abortion from the training and test set[6] released in the occasion of Stance Detection shared task organized by the authors of [10] in SemEval 2016. Each tweet is annotated with "against", "favor" and "none" labels toward the pre-chosen target. Hereafter we will refer to corpus involving feminist movement target as "Feminism" and corpus concerning legalization of abortion target as "Abortion".

Table 1 shows the composition of the two considered datasets.

3.1 Analysis of Corpora

For the scope of this study, we investigated the presence of offensive words in the training sets of the considered datasets. At this purpose, we carried out an analysis of corpora. In particular, we calculated the size of corpora and vocabulary, lexical richness and the number of offensive words. The lexical richness is calculated by means of the Type-Token Ratio (TTR) that calculates the variation of the lexicon into each corpus.

To individuate offensive words in the texts, we used various resources: English lexicons created by [6] about sexuality, human body, femininity and profanities; the NoSwearing English lexicon of swear words[7]; and the English version of Hurtlex[8] [2]. This resource, used for misogyny detection in [12], is divided in "conservative" and "inclusive" negative expressions.

These resources are described in Table 2.

In order to obtain these values, every symbol and punctuation is cleaned off as well as the urls. Considering the important role played by hashtags[9] and mentions (@user) in the tweet context, they are taken into account as tokens. The analysis is carried out considering the tweets labeled as "favor" and "against".

Table 3 and Table 4 summarize these aspects for each corpus.

As the tables show, the tweets annotated as "against" contain more offensive words than the favorable tweets, especially on the Feminist corpus.

Considering these premises we approached the stance classification taking into account the presence of hate speech.

4 APPROACH

On the basis of the previous observations, we implemented two dedicated systems able to classify the stance of the author's of the tweets toward respectively the abortion and feminism targets. We employed a classical machine learning approach guided by stylistic features and bigrams of words to detect the stance toward legalization of abortion, and lexical and semantic features for feminism target. In particular, we used a simple Support Vector Machine (SVM) classifier with radial basis function kernel (RBF) using the following parameters: $C = 5$ and $\gamma = 0.1$.

Considering the imbalanced collection of data (see Table 1), we used the function to balance the weights of the classes provided by Scikit-learn library[10]. Moreover, considering the multi-class ("against", "favor", "none") classification problem, to detect the correct class for each tweet, we predicted the probabilities of a tweet to belong to each class and then, we chose the class with the highest probability.

In order to evaluate the performances of our systems we compared the obtained results with the values obtained by the participating systems at Stance Detection shared task and the baselines provided by the organizers [10].

In the next sections, we describe in details the implemented features for both issues and the used evaluation metrics.

5 FEATURE ENGINEERING AND EXPERIMENTS

As said before, we implemented dedicated approaches for each of the two analyzed targets: legalization of abortion and feminist movement.

5.1 Legalization of abortion

Despite the analyses of the "Abortion" corpus showed the presence of hateful expressions in the texts annotated as "against" (see Table 3), the lexicons described in Table 2 do not help the classifier (see Section 6 and Table 7).

Therefore, we tried to capture relevant expressions by means of bigrams of words. In fact, analyzing in the training set the most relevant co-occurrences weighted with the Mutual Information measure, we noticed that the bigrams are enough informative and, in the majority of cases, also aggressive, such as: "dead woman", "human right", "right choose", "killing baby", "let live", "use someone", "death penalty", "black life". On the basis of this analysis, we extracted for each tweet bigrams of words lemmatized using WordNet lemmatizer provided by NLTK (Natural Language Toolkit) [11].

In addition, to consider stylistic nuances between favor and against labelled tweets, we captured sequences of characters from 3 to 5 grams. Both characters and words n-grams are weighted with the TF-IDF (Term Frequency-Inverse Document Frequency) measure.

In order to obtain informative grams, we pre-processed the texts deleting the url, the emoticons/emoji, the simbols of # and @, the abbreviation "RT" of a retweet, numbers and punctuation. As confirmation of our intuition, the bigrams improved the performance of the classifier of almost 3% compared to the simple use of characters n-grams (see Table 5).

5.2 Feminist movement

Differently from the former system, we improved the performance of the stance classifier toward the feminist movement using especially lexicons about abusive language and semantic features based on the similarity measure (see Table 6). In this classification, we used also stylistic features extracted by characters n-grams (from 3 to 5 grams) and unigrams of words weighted with the TF-IDF measure.

[6]The entire dataset is available online at https://saifmohammad.com/WebPages/StanceDataset.htm

[7]Demo online: https://www.noswearing.com/dictionary

[8]Hurtlex multilingual resource is available online: http://hatespeech.di.unito.it/resources.html

[9]The query hashtags used to extract tweets from Twitter are deleted by the organizers during the creation of the dataset to exclude obvious cues for the classification [10].

[10]https://scikit-learn.org/stable/

[11]https://www.nltk.org/

Table 1: Composition of the datasets.

	Training set			Test set		
	Favor	Against	None	Favor	Against	None
Feminism	210	328	126	58	183	44
Abortion	121	355	177	46	189	45

Table 2: Composition of the English lexicons.

Lexicon	*Num. of Words*	*Definition*
`Sexuality`	290	contains words related to sexuality such as $orgasm$, $anal$, $pussy$
`Human body`	50	contains words referred especially to feminine body such as ass, $vagina$ or $boobs$
`Femininity`	90	is a list of terms referring to women such as she, $girl$, $barbie$, $bitch$
`Offenses`	170	is a collection of vulgar words such as $pathetic$, $slut$ and $ugly$
`NoSwearing lexicon`	348	is a collection of swear words such as $whore$, $shitass$ and $buttfucker$
`Hurtlex conservative`	3953	contains common offenses such as *stupid*, *bitch*, *idiot*
`Hurtlex inclusive`	10175	includes words whose meanings are not hateful but in some contexts they could be used as offenses such as *barbarous*, *criminal*, *animal*

Table 3: Analysis of the "Abortion" corpus.

	Favor	Against
Number of tokens	1961	6052
Vocabulary	678	1739
Type-token ratio	34.57%	28.73%
Sexuality	113	291
Human body	20	68
Femininity	186	380
Offenses	9	26
NoSwearing lexicon	55	127
Hurtlex conservative	221	540
Hurtlex inclusive	524	1604

Table 4: Analysis of the Feminist corpus.

	Favor	Against
Number of tokens	3948	6027
Vocabulary	1313	1881
Type-token ratio	33.25%	31.21%
Sexuality	202	308
Human body	26	28
Femininity	410	684
Offenses	66	120
NoSwearing lexicon	116	201
Hurtlex conservative	432	773
Hurtlex inclusive	1029	1696

Carrying out the analysis of unigrams of words extracted in "Feminism" training set and weighting them with the Mutual Information measure, we noticed that, among the most relevant unigrams, there are various hashtags typical used to attack women or to defend them, such as: "yesallwomen", "spankafeminist", "feminazi", "weneedfemin" and "womensright". Therefore, even though we did not approach specifically the hashtags among our features, with unigrams the system is able to capture them.

About lexicons, we calculated for each tweet the number of words contained in each lexicon. The considered lexicons are the ones with higher statistical values in Table 4: lexicons about sexuality and femininity, NoSwearing lexicon and both Hurtlex lexicons.

Another important feature for this task is the similarity between the types of each tweet and the vocabularies of misogynistic and sexist collection of tweets. The former vocabulary was extracted by the tweets annotated as misogynistic in the English datasets released by the organizers of AMI shared tasks at IberEval and EvalIta 2018 [5]. The latter is extracted by the tweets annotated as sexist in the collection of tweets released by the authors of [15].

The similarity was calculate by cosine of similarity based on the word-embedding created on the base of these datasets. The word-embedding was created taking into account a window of 5 words and building a vector with a length of 100 items. The application of the lexical and semantic features about misogynistic and sexist speech improves the classifier performance of 7% compared to the simple use of character n-grams, overcoming all the baselines provided by the organizers of the competition (see Table 6).

6 EVALUATION AND RESULTS

For the evaluation, we compared the performances of our models with the results obtained by the participating systems at Stance Detection shared task. Therefore, we used the same measures of evaluation used in the competition. In particular, to evaluate the classification on "against" and "favor" predictions, they considered the f-scores for "favor" and "against" class calculated on precision and recall values.

The values of each measure was calculated as follows:

$$precision_{class} = \frac{correct\_class}{assigned\_class} \quad (1)$$

$$recall_{class} = \frac{correct\_class}{total\_class} \quad (2)$$

$$F_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}} \quad (3)$$

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \quad (4)$$

For the ranking, they used the average (f-avg) between the f-scores of "against" and "favor" classes. By taking into account the f-avg score, the "none" class is considered as negative class of the "against" and "favor" classes.

For proving the good performances of our models, we compared our results with the baselines provided by the organizers [10] and the result of the first rank system in the competition for each target.

These comparisons are showed in Table 5 and Table 6. In these tables, we report also the f-avg reached adding each feature.

Table 5: Results of stance detection toward the legalization of abortion

	f-avg
Baselines	
Majority class	40.30
SVM-unigrams	60.09
SVM-ngrams	*66.42*
SVM-ngrams-comb	63.71
First rank	63.32
Our Approach	
char-ngrams	66.30
+bigrams-words	**68.48**

Table 6: Results of stance detection toward the feminist movement

	f-avg
Baselines	
Majority class	39.10
SVM-unigrams	55.65
SVM-ngrams	*57.46*
SVM-ngrams-comb	52.82
First rank	**62.09**
Our Approach	
char-ngrams	52.65
+unigrams-words	53.83
+lexicons	58.06
+similarity	*60.27*

For the purpose of our work, precision and recall are optimal evaluation measures able to estimate the effectiveness of the implemented systems. To observe deeply the improvement reached in the classification, we analyzed the values of precision and recall obtained by each feature.

The precision evaluates how well the systems classify only the relevant documents. The recall reports the sensitivity of the model estimating how well the systems identify all relevant documents [3].

Table 7 and Table 8 show the increasing values obtained with each feature. In Table 7 we also report the values of recall and precision obtained adding lexical features related to abusive language and semantic information, which are not useful when we consider the "Abortion" stance target.

Moreover, we reproduced and reported in Table 7 and Table 8 the values of the most challenging baseline (the SVM-ngrams) as described on [10] [12]

In Table 7, we reported the performances of lexical features related to abusive and sexist language (Model 1) and semantic features (Model 2) for stance detection on the "Abortion" target compared with the used model (Our model) and the most challenging baseline (Baseline). As we can observe, abusive language extracted by means of lexicons achieved a higher value of recall for the Against class than our model. This means that in a real context where the social platforms or Internet companies want to retrieve almost all the possible unfavorable and offensive tweets, this feature could prove to be useful. However, the specialists should go through the false positives (i.e. the tweets predicted as against but actually favorable). In our case, the purpose is to find a balanced model, tuned on precision and recall, that is able to predict correctly both classes (Against and Favor).

Similarly to Model 1, the Model 2 using the cosine of similarity between the analyzed tweets and misogynistic and sexist tweets performed a higher recall for the Against class than our model. Although Model 2 seems to be more balanced than the previous one, the low recall for Favor class suggests that is not really able to retrieve favorable opinions.

As showed by the f-score values, our model seems to perform a more balanced prediction of both classes. In addition, compared to baseline' values, recall and precision of used model reached an overall increase differently from Model 1 and Model 2.

In Table 8, we reported the values of recall, precision and f-score obtained with the proposed model for "Feminism" stance detection compared with the baseline' values. Despite the recall for Favor class and the precision for the Against class are lower, the precision for Favor class and the recall for the Against class are higher than baseline' ones. Observing these results, our model, actually, seems to make more sensitive the system to retrieve unfavorable opinions than the baseline model. However, it achieves more balanced values of recall and precision for both classes than baseline model. In addition, compared to initial values obtained with characters n-grams and unigrams of words, we can observe that lexical and semantic features related to misogynistic and sexist speech guided the system to retrieve favorable and unfavorable tweets correctly.

These observations suggest that the used models are able to lead the classifiers to satisfy a balanced performance in precision and recall in both classes in accordance with the purposes of our investigation.

[12] The results of f-avg are slightly different from the ones reported in [10] (64.07 for "Abortion" and 55.52 for "Feminism" stance detection), but the obtained values of recall and precision are still significant for our analysis.

Table 7: Evaluation metrics for "Abortion" stance detection

	precision	recall	f-score
Baseline			
SVM-ngrams			
Favor	46.58	73.91	57.14
Against	80.54	63.49	71.01
Our model			
char-ngrams			
Favor	50.00	73.91	59.65
Against	82.12	65.61	72.94
+bigrams-words			
Favor	50.72	**76.09**	60.87
Against	**83.54**	69.84	76.08
Model 1			
char-ngrams+bigrams-words+lexicons			
Favor	48.15	28.26	35.62
Against	72.38	**80.42**	76.19
Model 2			
char-ngrams+bigrams-words+similarity			
Favor	**51.92**	58.70	55.10
Against	79.44	75.66	77.51

Table 8: Evaluation metrics for "Feminism" stance detection

	precision	recall	f-score
Baseline			
SVM-ngrams			
Favor	33.90	**68.97**	45.45
Against	**79.69**	55.74	65.59
Our model			
char n-grams			
Favor	32.29	53.45	40.26
Against	74.13	57.92	65.03
+unigrams-words			
Favor	32.98	53.45	40.79
Against	72.90	61.75	66.86
+lexicons			
Favor	38.04	60.34	46.67
Against	76.82	63.39	69.46
+similarity			
Favor	**41.57**	63.79	50.34
Against	76.28	**65.03**	70.21

7 DISCUSSION

To understand better where our systems fail the correct classification of the tweets, we carried out an error analysis. Looking at the misclassified tweets in both classifications, we noticed that some tweets miss the context and sometimes the only useful cue that suggest if the text is favorable or not is the hashtag. As underlined above, some hashtags used as queries to extract tweets were replaced with "#SemST" to exclude obvious cues for the classification [10], making this task more difficult even for humans. Below, we report some examples:

> (5) Some men do not deserve to be called gentlemen #SemST[13]

> (6) A much needed 3 days with these guys @rory3burke @Im_Brady missed @JimmahTwittah but what a weekend #SemST[14]

> (7) In civilian clothes and someone laughs at me thinking its a joke that I'm apart of the U.S. Navy. #SemST[15]

Other ones are hard to understand even finding the original hashtag, such as:

> (8) As I rewatced Charmed episodes! LOVING IT EVEN MORE! #SemST
> (8) *As I rewatced Charmed episodes! LOVING IT EVEN MORE! #feminism*

> (9) ..Can I also add that I really enjoyed looking at @TahirRajBhasin in #Mardaani :P Tahir, you were a dashing baddie! #Bollywood #SemST
> (9) *..Can I also add that I really enjoyed looking at @TahirRajBhasin in #Mardaani :P Tahir, you were a*

[13]The original tweet is: *Some men do not deserve to be called gentlemen #WomenAgainstFeminism.*
[14]The original tweet misses.
[15]The original tweet misses.

dashing baddie! #Bollywood #Feminism

In fact, tweets (8) and (9) refer to some specific and particular contexts far from the common world knowledge.

Some misclassified tweets contain irony like (10) and (11):

> (10) Equality is the police burying a domestic violence accusation against a female sports star, too #wedidit #usa #SemST[16]

> (11) @LifeSite Right, where are the pre-born women's rights? #allLivesMatter #equalRights #SemST[17]

Others are not specifically unfavorable toward the targeted issues but their gold annotation is "against", like (12) and (13):

> (12) Should start a "menism" movement. The amount of times people say "you've got tidy handwriting for a guy" is ridiculous #SemST[18]

> (13) Those who wonder what they would have done had they lived at the time of some terrible injustice now know the answer -P. Hitchens #SemST[19]

Considering the results that we obtained by our analyses, we think that dedicated approaches to detect stance toward some specific political and social issues could accentuate the sensitivity and the accuracy of the system.

Moreover, analyzing the correct and incorrect predicted tweets, we noticed that the majority of correct opposite tweets contains hateful expressions as a confirmation of our initial intuition. Thanks to the used features, our systems are able to classify correctly also the opinions that are more aggressive and offensive. Therefore, the proposed models are also able to deal with the identification of hate speech.

Below, we report some of these cases extracted from the correct predicted unfavorable tweets of both classifications:

> (14) I am about to deck these 2 bitches in the fucking mouth. #1A #2A #NRA #COS #CCOT #TGDN #PJNET #WAKEUPAMERICA #SemST

> (14) Meanwhile, @JustinTrudeau wants to waste your money to kill innocent children in the womb. #dangerous #hypocrite #noChoice #SemST

> (15) Women are taught to put their values into their hymens, rather than their intelligence, accomplishments, goals or character #feminism #SemST

> (16) You should start using Google translate @baedontcare, it is sooooo easy even retarded feminists like you can use it. #SemST

[16]The original tweet misses.

[17]The original tweet is: *@LifeSite Right, where are the pre-born women's rights? #prolife #allLivesMatter #equalRights*

[18]The original tweet misses.

[19]The original tweet misses.

8 CONCLUSION

The analyses carried out in this work show that the stance of users, especially unfavorable toward a pre-chosen target, often contain aggressive expressions aiming to offend and hurt the counterpart. This kind of expressions surely cannot guarantee constructive debates; on the contrary they incite and encourage misbehaviours that could have violent effects also in real life.

Despite this is a initial work, the obtained results confirm that it is possible to deal with hate speech identification in a stance detection process even with a simple model. With these premises and inspired by [16], we will extend our analyses on long texts online, such as news papers targeting the same pre-chosen target, and investigate the usefulness of the multi-learning approach. Furthermore, it would be interesting proving our findings on other social and political issues, such as immigration laws.

ACKNOWLEDGMENTS

The work of Simona Frenda was done during the six-months internship at the National Institute of Informatics in Tokyo, under the NII International Internship Program and supported by JSPS KAKENHI Grant Numbers JP18H03338, JP19H04420 and JPJP16H01756. The work of Viviana Patti was partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01). The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

REFERENCES

[1] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018*. Springer International Publishing, Paris, France, 57–64.

[2] Elisa Bassignana, Valerio Basile, and Patti Viviana. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of CLiC-it 2018 (CEUR Workshop Proceedings)*, Vol. 2253. CEUR-WS.org, Turin, 6.

[3] David C. Blair and M. E. Maron. 1985. An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. *Commun. ACM* 28, 3 (1985), 289–299.

[4] Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018. Overview of the Task on Automatic Misogyny Identification at IBEREVAL. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), September 2018*. CEUR-WS.org, Seville, Spain, 214–228.

[5] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Workshop EVALITA 2018*, Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso (Eds.). CEUR.org, Turin, Italy, 9.

[6] Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, and Luis Villaseñor Pineda. 2018. Automatic Expansion of Lexicons for Multilingual Misogyny Detection. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), December 12-13, 2018. (CEUR Workshop Proceedings)*, Vol. 2263. CEUR-WS.org, Turin, Italy, 6.

[7] Mirko Lai, Alessandra Teresa Cignarella, and Delia Irazú Hernandez Fariás. 2017. iTACOS at IberEval2017:Detecting Stance in Catalan and Spanish Tweets. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL),CEUR Workshop Proceedings. CEUR-WS.org, 2017* (19) *(CEUR Workshop Proceedings)*, Vol. 1881. CEUR-WS.org, Murcia, Spain, 185–192.

[8] Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2017. Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. In *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I*, Grigori Sidorov and Oscar Herrera-Alcántara (Eds.). Springer International Publishing, Cham, 155–168. https://doi.org/10.1007/978-3-319-62434-1_13

[9] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2017. Extracting Graph Topological Information and Users' Opinion. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings*, Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 112–118. https://doi.org/10.1007/978-3-319-65813-1_10

[10] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. ACL, San Diego, California, 31–41.

[11] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Trans. Internet Technol.* 17, 3, Article 26 (June 2017), 23 pages. https://doi.org/10.1145/3003433

[12] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) (CEUR Workshop Proceedings)*, Vol. 2263. CEUR-WS.org, Turin, Italy, 1–6.

[13] Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, Washington, US, 153–160.

[14] Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (AFNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 226–234.

[15] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. ACL, San Diego, California, 88–93.

[16] Masaharu Yoshioka, Myungha Jang, James Allan, and Noriko Kando. 2018. Visualizing Polarity-based Stances of News Websites. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), March 26, 2018. (CEUR Workshop Proceedings)*, Vol. 2079. CEUR-WS.org, Grenoble, France, 6–8.