Merging datasets for aggressive text identification

Paula Fortuna[1] **José Ferreira**[1,2] **Luiz Pires**[3] **Guilherme Routar**[2] **Sérgio Nunes**[1,2]

(1) INESC TEC and (2) FEUP, University of Porto and (3) FCUP, University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
paula.fortuna@fe.up.pt, sergio.nunes@fe.up.pt

Abstract

This paper presents the approach of the team "groutar" to the shared task on Aggression Identification, considering the test sets in English, both from Facebook and general Social Media. This experiment aims to test the effect of merging new datasets in the performance of classification models. We followed a standard machine learning approach with training, validation, and testing phases, and considered features such as part-of-speech, frequencies of insults, punctuation, sentiment, and capitalization. In terms of algorithms, we experimented with Boosted Logistic Regression, Multi-Layer Perceptron, Parallel Random Forest and eXtreme Gradient Boosting. One question appearing was how to merge datasets using different classification systems (e.g. aggression vs. toxicity). Other issue concerns the possibility to generalize models and apply them to data from different social networks. Regarding these, we merged two datasets, and the results showed that training with similar data is an advantage in the classification of social networks data. However, adding data from different platforms, allowed slightly better results in both Facebook and Social Media, indicating that more generalized models can be an advantage.

1 Introduction

In the last few years, we have witnessed a growing number of online platforms where users can post content. As the number of platforms has increased, so has the number of aggressive interactions, such as cyberbullying or hate speech. The goal of our work is to contribute to the automatic identification of this type of communication through the participation in the Shared Task on Aggression Identification in text (Kumar et al., 2018a).

The task consisted in developing a classifier that could make a 3-way classification between Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-aggressive (NAG) text data. The organizers provided a dataset of 15,000 aggression-annotated Facebook posts for training and validating the classification systems. Each team was allowed to test up to three systems and to use additional data for training, as long as the data would be publicly available before submission of the system paper. Different competitions were available with variations in language and data sources. It was possible to classify aggression in English and Hindi and also using data from Facebook or other Social Media platform unknown at submission time.

In our approach, we focused on understanding the effects of merging new datasets for training models. We used the Toxicity dataset, from the Toxic comment classification challenge in Kaggle, as an additional source of data, and we proceeded with the conversion from toxicity to aggression. We built and compared two systems, one using only the original data for training, and the second using also toxic data. We extracted some classic features and studied different machine learning classification algorithms using a methodology of training, validation, and testing. Our approach focused on English and in both test sets provided, Facebook and Social Media.

In the next sections, we present the related work (Section 2), our method (Section 3), the results (Section 4), and finally our conclusions (Section 5).

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 128–139
Santa Fe, USA, August 25, 2018.

2 Related Work

Aggression by text is a complex phenomenon, and different knowledge fields try to study and tackle this problem. In this analysis of related work, we focus mainly on a computer science perspective on aggression identification, a recent emerging area. In the last years, the scientific study of automatic identification of aggressive text, from a Computer Science and Engineering point of view, is increasing. One important aspect to consider is that in this scientific community, several related terms are used to express different types of aggression. Some of those are hate (Tarasova, 2016), cyberbullying (Chen, 2011), abusive language (Nobata et al., 2016), profanity (Dictionary, 2017), toxicity (Jigsaw, 2017), flaming (Guermazi et al., 2007), extremism (Prentice et al., 2011; McNamee et al., 2010), radicalization (Agarwal and Sureka, 2015), and hate speech (Schmidt and Wiegand, 2017).

Despite the differences between those concepts, previous research can give us insight into how to approach the problem of identifying aggressive interactions. For instance, particular attention has been given to the automatic detection of hate speech. In one survey paper (Schmidt and Wiegand, 2017), the authors provide a short, comprehensive, structured and critical overview of the field of automatic hate speech detection in natural language processing. In other, the main focus is on definitions and rules for classification (Fortuna and Nunes, forthcoming), which is important for solving this complex task. One of the main conclusions of these works is that the automatic classification of hate speech and other related concepts rely frequently upon Machine Learning and classification approaches.

Regarding the automatic classification of messages, one first step is the gathering of training data. Several studies published datasets for aggression identification with different classification systems. For example, in one dataset the classes "Racism", "Sexism" and "Neither" were used to annotate tweets for English (Waseem and Hovy, 2016). In another dataset collected for the specific topic of hate speech against refugees, tweets in German were annotated using only the class "Hate Speech" (Ross et al., 2017). Another study presents a hate speech detection dataset in Twitter for English, using the classes "Hate", "Offensive" or "Neither" (Davidson et al., 2017). A third dataset not publicly available contains comments from Yahoo! Finance and News in English and uses the classes "Hate Speech", "Derogatory", "Profanity" and "Neither" (Nobata et al., 2016). Finally, one last dataset from a classification challenge in Kaggle identifies Toxicity (Jigsaw, 2018). The dataset contains Wikipedia comments marked as "toxic", "severe toxic", "obscene", "threat" and "identity hate", in a multi-class and multi-label approach. Based on the information from these datasets, we conclude that none considers the class "aggression", which would be useful for this work. Another difficulty is the multiplicity of different concepts and definitions. A recent work identifies this problem and proposes a typology that captures the similarities between concepts (Waseem et al., 2017). According to this typology, abuse follows into directed vs. generalized and explicit vs. implicit categories. This topology has implications on the following parts of a classification procedure.

After the data collection, one of the most important steps when using classification is the process of feature extraction (Schmidt and Wiegand, 2017). Different approaches are being used, ranging from Dictionaries (Liu and Forss, 2015; Dadvar et al., 2012; Dinakar et al., 2011), to Bag-of-words (Burnap and Williams, 2016; Kwok and Wang, 2013; Greevy and Smeaton, 2004), N-grams (Burnap and Williams, 2016; Nobata et al., 2016; Waseem and Hovy, 2016; Liu and Forss, 2014; Greevy and Smeaton, 2004; Badjatiya et al., 2017; Davidson et al., 2017), Part-of-speech (Greevy and Smeaton, 2004; Dinakar et al., 2011; Burnap and Williams, 2014), Lexical Syntactic Feature-based (LSF) (Chen, 2011), Rule based approaches (Haralambous and Lenca, 2014), Participant-vocabulary consistency (PVC) (Raisi and Huang, 2016), Template-based Strategies (Powers, 2011), Word Sense Disambiguation Techniques (Yarowsky, 1994), Sentiment analysis (Liu and Forss, 2014; Liu and Forss, 2015; Gitari et al., 2015; Agarwal and Sureka, 2017; Del Vigna et al., 2017; Schmidt and Wiegand, 2017; Davidson et al., 2017), perpetrator characteristics (Waseem and Hovy, 2016), Paragraph2vec (Djuric et al., 2015) and Deep learning (Yuan et al., 2016). There are also features and approaches more specific to the problem of hate speech detection, namely: othering language (Burnap and Williams, 2016; Dashti et al., 2015), declarations of superiority of the ingroup (Warner and Hirschberg, 2012), objectivity (Gitari et al., 2015) and subjectivity (Warner and Hirschberg, 2012) of hate speech language. Additionally, in the typology of hate,

some considerations are made regarding the features to use (Waseem et al., 2017). In the case of direct abuse, mentions, proper nouns, named entities, and co-reference resolution can be helpful. In generalized abuse, researchers should consider identifying vocabulary specificities regarding the groups targeted. On the other hand, explicit abuse is often indicated by specific keywords. Hence, dictionary-based approaches may work well. Finally, implicit abuse identification works with character N-grams (Mehdad and Tetreault, 2016), word embeddings (Djuric et al., 2015) and perpetrator characteristics (Waseem and Hovy, 2016).

Regarding the classification algorithms, the more common are SVM (Del Vigna et al., 2017), Random forests (Burnap and Williams, 2014), Decision trees (Dinakar et al., 2011), Logistic regression (Davidson et al., 2017), Naive bayes (Liu and Forss, 2015) and Deep learning (Yuan et al., 2016).

In this challenge, we are not only interested in distinguishing between aggressive and non-aggressive text, but different degrees of aggression are also considered. A recent discussion on the challenges of identifying profanity vs. hate speech highlighted some issues in this topic (Malmasi and Zampieri, 2018). The results revealed that discriminating hate speech from profanity is not a simple task, and it may require features more complex than N-grams. From this conclusion, we can extrapolate that distinguishing overt and covert aggression will be difficult as well. Overcoming this difficulty is a motivating factor for conducting this shared task.

Regarding the specificities of our approach, the main research question of our work concerns the effects of merging new datasets on the performance of models for aggression classification. Additionally, there are some open issues that motivate our work. One question is if it is possible the combination of datasets annotated with different classification systems (e.g. toxicity and aggression). This combination would allow the use of multiple datasets simultaneously. Another question is if it is possible to generalize models and apply them to data from different Internet sources. Finally, other question concerns the duration in time of the models, even when the same platform is used, due to the fast evolution of online language.

In the next sections, we aim to answer to some of these questions with our approach.

3 Methodology and Data

3.1 The datasets

The provided training datasets (Kumar et al., 2018b) contained Facebook text messages for English and Hindi. From those messages, 12,000 were for training and 3,000 for development (dev). This last was a dataset for testing before submitting final results. Regarding the test set, was the data available for final classification and final submission for ranking of solutions in the contest. Several scenarios were available for testing the final models. Besides different languages (English and Hindi), the teams could classify diverse message sources (Facebook and general Social Media). For the annotation of the datasets, there were three classes described solely as Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-aggressive (NAG) and no additional information provided. This lack of deeper definitions opposes to previous recommendations (Ross et al., 2017), which pointed out the importance to clearly define concepts before addressing problems like hate speech identification.

Aiming to improve the available definitions, we tried to manually inspect some data, so that we could better understand the differences among the three types of messages. We concluded that it is not easy to distinguish the classes and that they overlap (Table **??**). For example, a text like "Nonsense" is marked as overtly aggressive (OAG), while "No respect for him now" is marked as non-aggressive (NAG).

Outside of the challenge, the definitions of the classes became available in the article presenting the dataset (Kumar et al., 2018b). According to these authors, overt aggression is any speech or text in which aggression is overtly expressed, either through the use of specific kind of lexical items, lexical features or certain syntactic structures, considered aggressive. On the other hand, covert aggression is any text in which aggression is not overtly expressed. It is an indirect attack against the victim and is often packaged as insincere polite expressions, through the use of conventionalised polite structures. For instance, cases of satire or rhetorical questions may be classified as covert aggression.

Considering the opportunities this task enabled, we decided to conduct our experiment only for English

Id	Text	Class
facebook_corpus_msr_326287	This is a false news Indian media is simply misguiding there nation and creating hatred.. Media should be v careful while spreading the news.. SHAME.:(	NAG
facebook_corpus_msr_1805657	No respect for him now	NAG
facebook_corpus_msr_401603	Now time has come to take firm action against pakistan, pl do not seat idle.....public anger....	NAG
facebook_corpus_msr_382223	Unfortunately this is wat indian govt is capable of doing!!!!...i dint vote for modiji to see such crap..	CAG
facebook_corpus_msr_470981	I visited 5 atm but I cont able to withdraw from money..not working..	CAG
facebook_corpus_msr_492174	I wanna meet the girl who said the iPhone is user friendly!!!	CAG
facebook_corpus_msr_1853672	What the hell is happening	OAG
facebook_corpus_msr_2032108	#salute you my friend	OAG
facebook_corpus_msr_2241597	Nonsense	OAG

Table 1: Examples of messages extracted from the provided dataset.

and to test the effect of adding a new dataset in our classification both in messages from Facebook and general Social Media. Despite our intention, we did not find an alternative dataset that would have classified text for aggression. We decided then to use a Toxicity dataset, already mentioned in Section 3.1. The Toxicity dataset consists of 170,355 messages marked as toxic, severe toxic, obscene, threat and identity hate, in a multi-class and multi-label approach (Jigsaw, 2018).

When we try to match the Aggression dataset with the Toxicity dataset, we are in the presence of two unequal classification systems. Therefore, we conducted a procedure for converting the classes of the Toxicity dataset into aggression (Figure 1).

Figure 1: Procedure conducted for transforming the Toxicity dataset into Aggressive communication dataset.

The steps followed are:

- Regarding the toxic column, we decided to ignore it because it correlates strongly with the others.
- The columns “severe toxic”, ‘insult” ,“obscene” and “threat” would correspond to “overtly aggressive” (OAG).
- The column “identity hate” would correspond to “covertly aggressive” (CAG).
- We excluded the instances that would score in both OAG and CAG dimensions because that is not possible in our original dataset.

In this procedure, we decided to only keep “identity hate” in the covertly aggressive (CAG) class. Following previous studies (Malmasi and Zampieri, 2018), “profanity vs. hate speech” are considered and both identified and handled as different classes. In the case of the Toxicity dataset, we think that

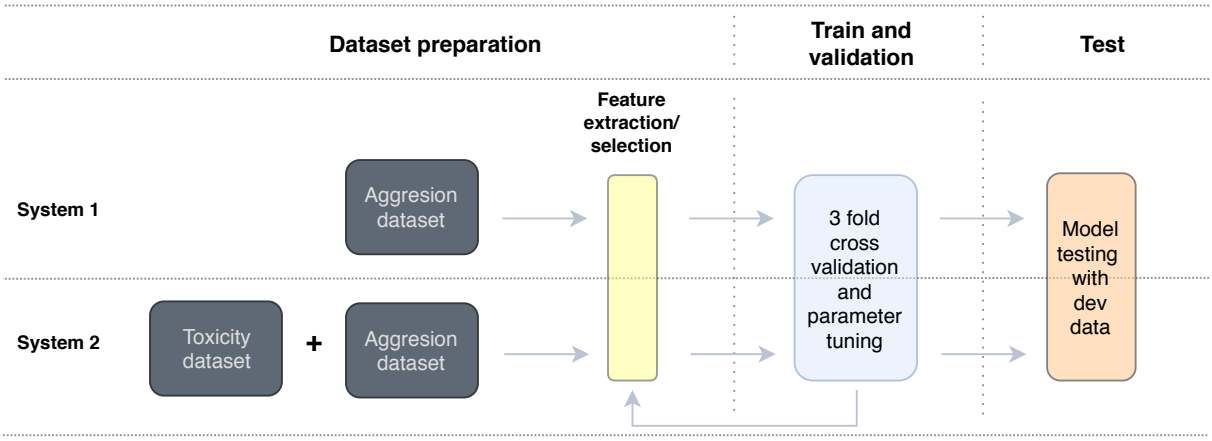


Figure 2: Conducted method for comparison of Systems 1 and 2 using training, validation, and test.

"severe toxic", "insult" ,"obscene" and "threat" are more similar to "profanity" than to "hate speech", and the four should be grouped together as overtly aggressive (OAG).

3.2 Method

In order to test the effect of adding a new dataset to our classification procedure, we compared two different systems (Figure 2). In the first, we build a model using only the training set provided in the contest. For that, we extracted some classic features and studied different machine learning classification methods. In the second, we applied the same procedure, but we fed the model not only with the provided data but also with the Toxicity dataset.

3.2.1 Dataset Preparation and Feature Extraction

Regarding the features, we used the NLTK 3.3 library (Bird et al., 2009) for extracting:

- Parts of speech (POS).
- Sentiment analysis.
- Combination of POS and sentiment analysis.
- Capitalized words.
- Punctuation patterns.
- Frequencies of insults.

The procedure consisted in tokenization and extraction of parts of speech (POS) with Penn Treebank style. Regarding sentiment, we considered Vader (Valence aware Dictionary and sentiment reasoner), a lexicon and rule-based sentiment analysis tool specialized in social media (Hutto and Gilbert, 2014). It produces four sentiment metrics, namely: positive, negative, neutral and compound. Additionally, to these metrics, we extracted the counts of negative words, and the counts of negative adjectives, combining both POS and sentiment analysis. We also measured the frequencies of capitalized words in a message and marked with a boolean full capitalized messages. Punctuation patterns were obtained as explained in Table **??**. Finally, we mapped the frequencies of insults, using a dictionary[1] with 350 words. The total number of features in each group is presented in Table **??**.

[1] http://www.insult.wiki/wiki/Insult_List

Individual features	Expression	Description
ellipsis	\.{2,}	ellipsis occurrence counts
ellipsis_reps	—	sum of the summed length of all ellipsis patterns
simple_qm	^?$	counts of single question mark
simple_exc	^!$	counts of single exclamation mark
reps_qm	^(\?+)$	counts of question marks with repetition
reps_exc	^(!+)$	counts of exclamation marks with repetition
mixed	(\?—\!){1,}	counts of patterns with both question and exclamation marks
num_punct	—	counts of punctuation patterns
max_punct	—	size of largest punctuation pattern

Table 2: Extracted features based on punctuation and regular expression used.

Feature group	Total features
Insult words	350
POS	36
Punctuation	9
Sentiment	5
Capitalization	2
POS + sentiment	2

Table 3: The total number of features by group.

3.2.2 Train and validation

In this phase, we used the R caret package (Kuhn, 2008) and the functions train, trainControl, predict and confusionMatrix. We opted by three-fold cross-validation with parameter tuning of length three. Regarding the classification algorithms, we used Boosted Logistic Regression (LogitBoost), Multi-Layer Perceptron (mlp), Parallel Random Forest (parRF) and eXtreme Gradient Boosting (xgbTree).

3.2.3 Test

For testing our model we conducted four different runs. We developed two systems (aggression data vs. aggression + toxicity data) that were tested in two different scenarios (Facebook data vs. Social Media). Based on the results of the train and validation phases, we submitted the following systems (Table **??**), for English data: training with the provided dataset, classification algorithm with parallel random forests and testing in Facebook data (Fb_ag_rf); training with the provided dataset plus the dataset classified on toxicity, classification algorithm with parallel random forests and testing in Facebook data (Fb_ag_tox_rf); training with the provided dataset, classification algorithm with parallel random forests and testing in Social Media data (Sm_ag_rf); training with the provided dataset plus the dataset classified on toxicity, classification algorithm with parallel random forests and testing in Social Media data (Sm_ag_tox_rf).

System id	Data for testing	Data for training	Classification Algorithm
Fb_ag_rf	Facebook	Aggression	Random forests
Fb_ag_tox_rf	Facebook	Aggression + Toxicity	Random forests
Sm_ag_rf	Social Media	Aggression	Random forests
Sm_ag_tox_rf	Social Media	Aggression + Toxicity	Random forests

Table 4: Systems considered in the submission, each corresponding to one run.

4 Results

4.1 Train and validation results

After training our models using cross-validation, and tuning them in the default parameters, we tested them in the development dataset. This section provides the results in this phase, which were used to decide on which classification algorithm to keep for the final submission. We concluded (Figure 3) that the models built using only the default dataset (marked as "without toxic dataset" in the figure) perform better than the ones using also the Toxicity dataset (marked as "with toxic dataset").

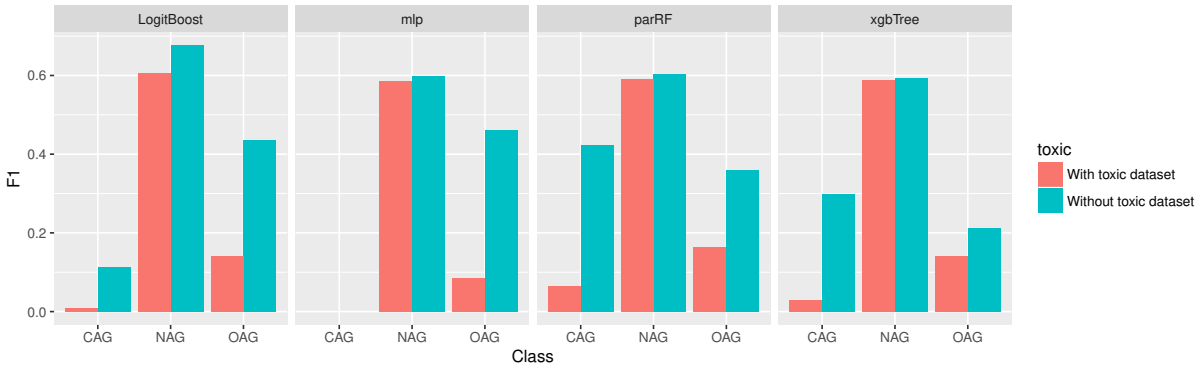


Figure 3: Results of different algorithms on the dev test set.

This is an expected result when using the development data for testing due to the source of its messages. In this case, the origin is Facebook, which is the same as in the training set. On the other side, the Toxicity dataset comes from a different platform (Wikipedia) and these extra messages can cause noise. Despite this result, we think that it worths to test the system built with the Toxicity dataset because on the final submission some test sets include unknown data. Regarding the classification algorithms, we decided to keep the parRF because it was the best performing algorithm if we take into account the results of the three classes.

4.2 Final test results

In this section, we present the results after submitting our classifications in the shared task platform. In Table 5, we can observe the results with the Facebook test set. With a mixed training dataset (Fb_ag_tox_rf system), we have a model with a slightly better performance than when using only the provided dataset for training (Fb_ag_rf system). In this case, we would expect the same results as in the development dataset because in both cases data originates in Facebook. Hence the model trained without Toxicity should have performed better. Regarding this unexpected result, one possible explanation would be if the moments for the collection of both dev and test set would not match, and therefore some differences existing due to that.

	CAG	**NAG**	**OAG**	**avg**
random baseline	-	-	-	0.3535
Fb_ag_rf	0.2135	0.6379	0.3439	0.5259
Fb_ag_tox_rf	**0.2217**	**0.6403**	**0.3439**	**0.5288**

Table 5: Results for the English (Facebook) task, comparing the use of aggressive data (FB_ag_rf) with the use of aggressive data plus Toxicity dataset (FB_ag_tox_rf) for training using Random Forests for classification.

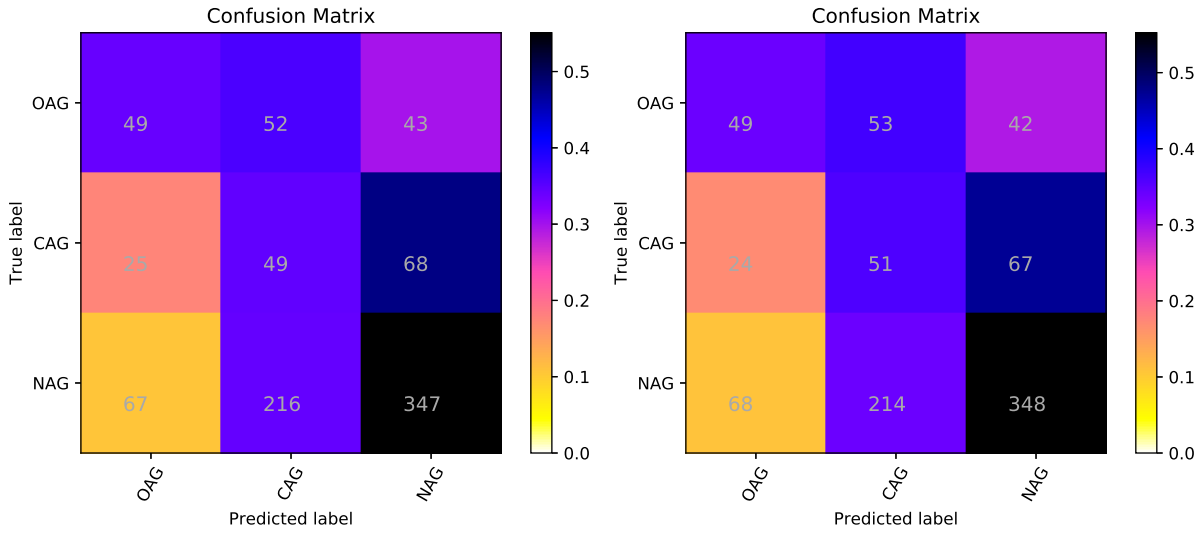
We verified the same pattern described from the Facebook test set on the Social Media data (Table 6). Using a mixed dataset for training lead to models with a better performance. In this case, this is an expected result because the Social Media messages are from another source than Facebook and therefore a more generic model is likely to perform better.

	CAG	NAG	OAG	avg
random baseline	-	-	-	0.3477
EN-TW task, groutar 00	0.314	0.4863	0.2469	0.3609
EN-TW task, groutar 01	**0.3151**	**0.4889**	**0.2505**	**0.3633**

Table 6: Results for the English (Social Media) task, comparing the use of aggressive data (Sm_ag_rf) with the use of aggressive data plus Toxicity dataset (Sm_ag_tox_rf) for training using Random Forests for classification.

If we compare both (Table 5 and Table 6), we achieved an overall better performance when classifying Facebook than Social Media messages. This is also an expected result because we trained with messages from this social network and the added Toxicity dataset originates in Wikipedia. On the other hand, this pattern does not apply if we consider only the covertly aggressive messages (CAG). In this case, the classification worked better in the Social Media messages. This supports the idea that different social media platforms have different expressions of behavior and the covertly aggressive messages were easier to target on Twitter.

We also present here the confusion matrix for the Facebook test set and the Social Media (Figures 4 and 5). We concluded that, when adding the Toxicity dataset, the results are slightly better in identifying covertly aggressive messages (CAG) on Facebook, overtly aggressive messages (OAG) on Social Media and non-aggressive messages (NAG) in both.



(a) Provided dataset

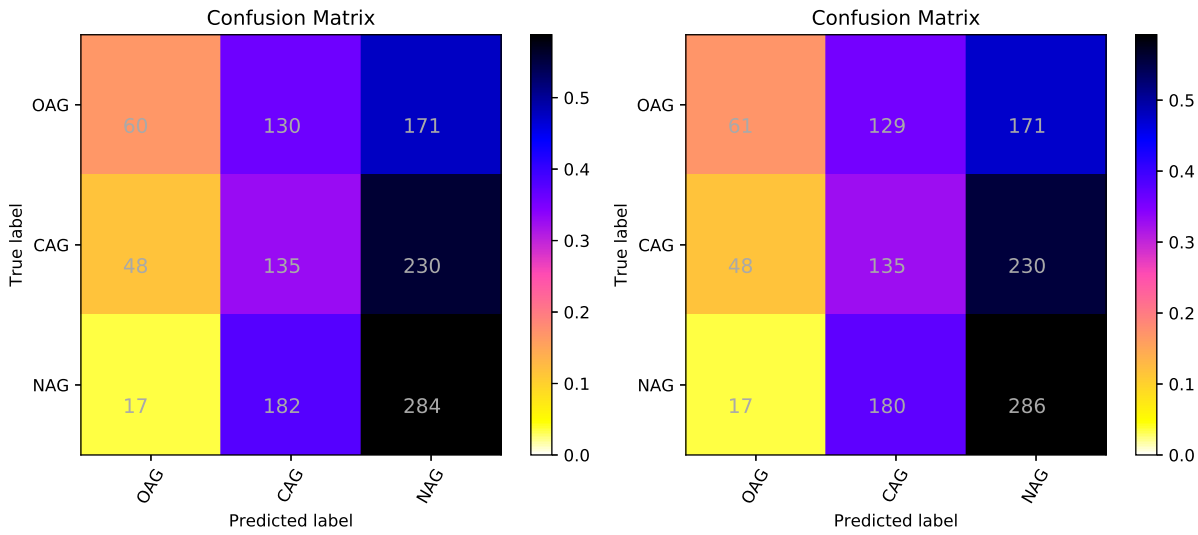
(b) Provided + Toxicity datasets

Figure 4: Confusion matrix for the two developed systems, using the Facebook test set and Random Forest classification algorithm.

5 Conclusion

Throughout our approach to this shared task, our goal was to discuss some open issues in aggressive text identification. Namely, the main motivation of our work was measuring the effects of merging new datasets on the performance of models for aggression classification. It can be difficult to combine distinct datasets due to the differences in the classification systems used. We conducted an experiment where we combined a toxicity dataset with the original aggression dataset used in this shared task. Our expectation was that, by adding external data from a different context, we could improve the performance of the system.

In the procedure of adding different datasets to the data from the shared task, we faced one problem.



(a) Provided dataset

(b) Provided + Toxicity datasets

Figure 5: Confusion matrix for the two developed systems, using the Social media test set and Random Forest classification algorithm.

We found no alternative dataset with text classified for aggression and therefore we had to merge datasets using different classes (aggression vs. toxicity). This required a conversion procedure where we corresponded identity hate with covertly aggressive discourse, and severe toxic, insult, obscene and threat, were mapped to overtly aggressive discourse.

Another goal of our work was focused on evaluating if we can build models that are general enough to be useful across different social networks. From our experiments, on average we achieved a better performance in classifying messages from the same social network that we used for training (Facebook) when comparing to other social media. This confirms that training with similar data is an advantage in the classification of social networks data. However, on the other hand, adding data for training from a different platform, allowed us to slightly increase performance, indicating that more generalized models can be an advantage.

Regarding the features, we used POS tags, sentiment analysis, insult frequencies, capitalization, and punctuation counts. According to the literature, this kind of features are more related with explicit abuse detection (Waseem et al., 2017). However, we did not find any evident advantage in using them for detecting overtly aggressive discourse (OAG) in comparison with covertly aggressive (CAG). In our experiment, the results of classifying OAG and CAG were equivalent. This can be due to the simplicity of the extracted features, or possibly to some weaknesses in the data, as we explain in the next paragraph.

In the exploration of the dataset, we faced unclear definitions of the classification system used in the annotation. Also, the definitions provided a posteriori seemed to be superficial. We manually inspected some messages and concluded that it was difficult to identify the differences between the classes because messages with similar degrees of aggression were found in the three classes. Additionally, we also found lack of clear definitions in the toxicity dataset. This problem should be tackled in future research because the identification of aggression is complex and ambiguous even for humans and requires clear guidelines. Finally, we also noticed a higher percentage of aggressive messages in this dataset in comparison to previous studies in other related phenomenon (Davidson et al., 2017), which questions the quality of the annotation.

Acknowledgements

This work was partially funded by FourEyes, a Research Line within project “TEC4Growth Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01- 0145-FEDER-000020”, financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL

2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Swati Agarwal and Ashish Sureka. 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, volume 43.

Peter Burnap and Matthew L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In *Proceedings of Internet, Policy & Politics*, pages 1–18.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Ying Chen. 2011. *Detecting Offensive Language in Social Medias for Protection of Adolescent Online Safety*. Ph.D. thesis, The Pennsylvania State University.

Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop*, pages 23–25. University of Ghent.

Ali A. Dashti, Ali A. Al-Kandari, and Hamed H. Al-Abdullah. 2015. The influence of sectarian and tribal discourse in newspapers readers' online comments about freedom of expression, censorship and national unity in kuwait. *Telematics and Informatics*, 32(2):245–253.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity*, pages 86–95.

Cambridge Dictionary. 2017. Profanity. Available in `https://dictionary.cambridge.org/dictionary/english/profanity`, accessed last time in June 2017.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02).

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM2.

Paula Fortuna and Sérgio Nunes. forthcoming. A survey on automatic detection of hate speech in text. *ACM computing surveys (CSUR)*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Edel Greevy and Alan F. Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM.

Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou. 2007. Using a semi-automatic keyword dictionary for improving violent web site filtering. In *Signal-Image Technologies and Internet-Based System, 2007. SITIS'07. Third International IEEE Conference on*, pages 337–344. IEEE.

Yannis Haralambous and Philippe Lenca. 2014. Text classification using association rules, dependency pruning and hyperonymization. *arXiv preprint arXiv:1407.7357*.

CJ J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and ...*, pages 216–225.

Jigsaw. 2017. Perspective api. Available in `https://www.perspectiveapi.com/` , accessed last time in June 2017.

Jigsaw. 2018. Toxic comment classification challenge identify and classify toxic online comments. Available in `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`, accessed last time in 23 May 2018.

Max et al. Kuhn. 2008. Caret package. *Journal of statistical software*, 28(5):1–26.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Association for the Advancement of Artificial Intelligence*.

Shuhua Liu and Thomas Forss. 2014. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 530–537.

Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, volume 1, pages 487–495. IEEE.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Lacy G McNamee, Brittany L Peterson, and Jorge Peña. 2010. A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, 77(2):257–280.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

David Martin Powers. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63.

Sheryl Prentice, Paul J Taylor, Paul Rayson, Andrew Hoskins, and Ben O'Loughlin. 2011. Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the gaza conflict. *Information Systems Frontiers*, 13(1):61–73.

Elaheh Raisi and Bert Huang. 2016. Cyberbullying identification using participant-vocabulary consistency. *arXiv preprint arXiv:1606.08084*.

Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. *SocialNLP 2017*, page 1.

Natalya Tarasova. 2016. Classification of hate tweets and their reasons using svm. Master's thesis, Uppsala Universitet.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 88–95. Association for Computational Linguistics.

Shuhan Yuan, Xintao Wu, and Yang Xiang. 2016. A two phase deep learning model for identifying discrimination from tweets. In *International Conference on Extending Database Technology*, pages 696–697.