A Hierarchically Labeled Portuguese Hate Speech Dataset

Paula Fortuna[1,3]
João Rocha da Silva[1,2]
Juan Soler-Company[3]
Leo Wanner[3,4]
Sérgio Nunes[1,2]

NEW CLASSIFICATION SCHEME

In this work, we propose that a more **fine-grained** view can be useful in hate speech classification, such that creating a language model for each category may be helpful to improve the automatic detection of hate speech [1]. Another phenomenon when analyzing different categories of hate speech is their ***intersectionality***:

- This concept brings attention to the experiences of people who are subjected to multiple forms of discrimination (e.g., being woman and black) [2].

We propose to use a **rooted DAG** in order to be able to cover hate speech subtypes and their intersections. For the annotation of this dataset, we defined two different schemes:

- First, non-experts annotated the tweets with binary labels.
- Then, expert annotators classified the tweets following a hierarchical multiple label scheme with 81 categories.

MESSAGE COLLECTION

We followed the steps:

- **Pages and keywords enumeration** - We looked at specific profiles, keywords and hashtags in a total of 58 search instances.
- **Crawling** - We used R to crawl a total of 42,930 tweets.
- **Filtering** - We filtered those messages and kept tweets written in Portuguese, eliminated repetitions, removed HTML tags and messages with less than three words.
- **Sampling** - We decided then to use a maximum of 200 tweets per search instance in order to keep a more diverse source of tweets.

Our final dataset contains 5,668 tweets, from 1,156 different users.

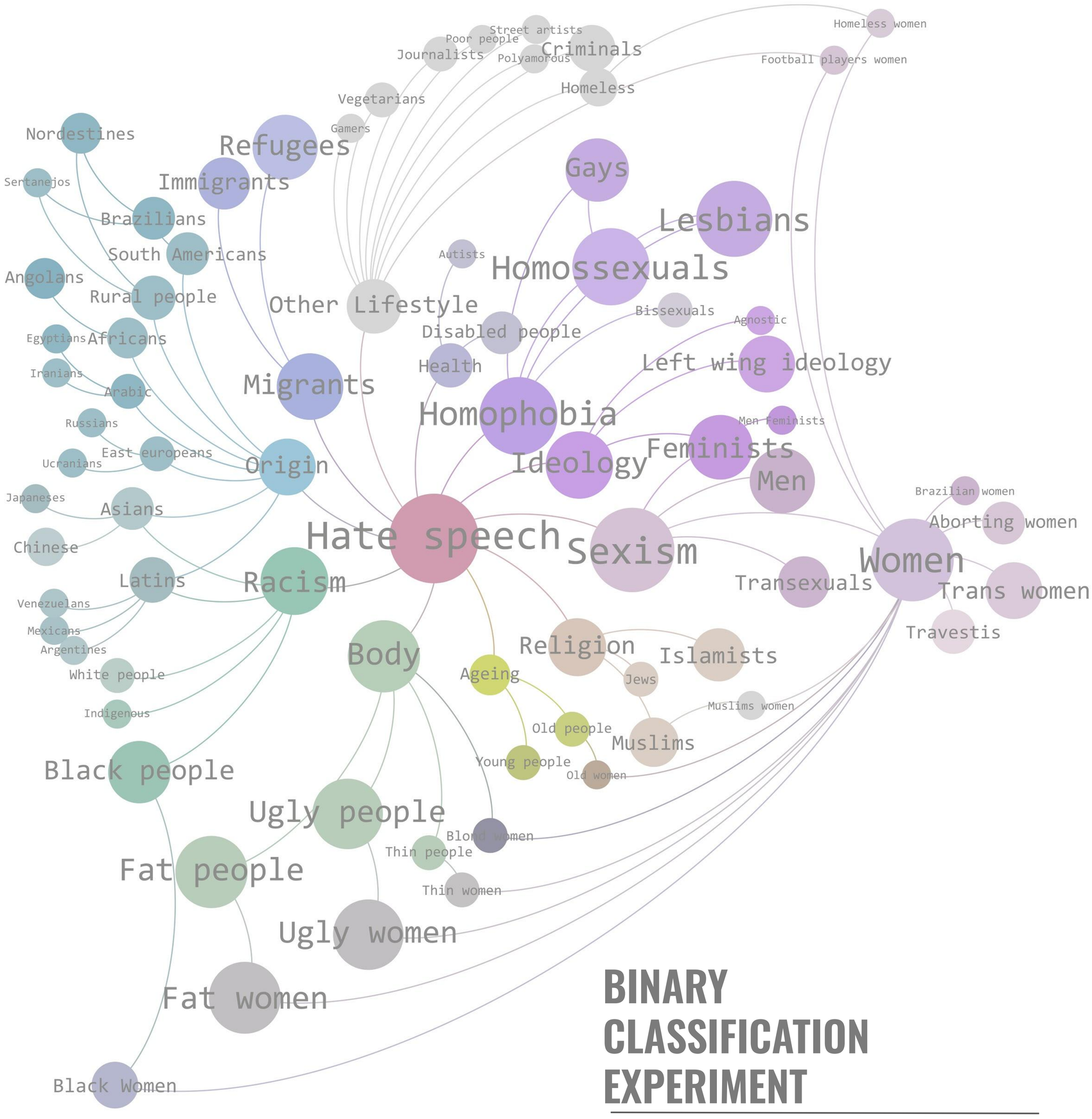
BINARY ANNOTATION

Three annotators classified every message. 18 Portuguese native speakers were given annotation guidelines and had to label each message as 'hate speech' or 'not hate speech'. We observed a low agreement with a Fleiss's Kappa [4] value of K = 0.17.

BUILDING THE HIERARCHY

Similarly to another work [5], we use for the annotation a **data-driven approach** based on an open coding methodology:

- The classification hierarchy is then built by creating and reorganizing categories until all available data is analyzed.
- We enumerate all the groups cited in our dataset, no matter their frequency.



The hierarchy of classes was built by one researcher with training in social psychology. For verifying the validity of this annotation procedure, a second annotator classified 500 messages. We observed an **annotator agreement** with Cohen's Kappa [6] value of K = 0.72. We also analyzed it by type of hate. We found diverse values in the different categories (Table 1), which points out that some specific types of hate speech can be more difficult to classify than others.

Class	K
Lesbians	0.88
Health	0.86
...	...
Gays	0.30
Ugly women	0.28

Table 1 - Annotator agreement per class.

BINARY CLASSIFICATION EXPERIMENT

For the experiement:

- We use 10-fold **cross-validation** with holdout validation.
- We remove stop words and punctuation.
- We use **pre-trained Glove word embeddings** with 300 dimensions for Portuguese [7].
- We use a deep learning model, namely **LSTM**, in an architecture as already proposed by [8].

Table 2 shows the achieved baseline results on our new dataset.

	New dataset
CV F1	0.78
Training data (N)	5099
Test set F1	0.72
Testing data (N)	567

Table 2 - Binary classification results

CONCLUSION

We provided a hate speech hierarchical labeling schema that integrates the complexity of hate speech subtypes and their intersections. This allowed us to find out that distinct types of hate speech present different agreement levels between annotators. Therefore, future guidelines for annotation may benefit from specifying the particularities of the different subtypes of hate speech. Finally, in future explorations of this dataset, we will experiment with multilabel classification of hate speech to identify not only whether a message contains hate, but also the targeted groups.

Acknowledgements

This work was partially funded by the Google DNI project Stop PropagHate. Soler-Company and Wanner have been supported by the European Commission under the contract numbers H2020--7000024-RIA and H2020-786731-RIA. We would like to thank the anonymous reviewers for their insightful comments and to the annotators for their contribution to this work.

References

[1] 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.
[2] Patricia Hill Collins. 2015. Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 41:1– 20.
[3] Hatebase. 2019. Hatebase. Available in https: //www.hatebase.org/, accessed last time in February 2019.
[4] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
[5] Joni Salminen, Hind Almerekhi, Milica Milenkovic´, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning mod- els for identifying and classifying hate in online news media. In T*welfth International AAAI Conference on Web and Social Media*.
[6] Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall's. 2012. Package Various coefficients of interrater reliability and agreement.
[7] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.
[8] 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.

The author's institutes are (1) INESC TEC (2) FEUP, University of Porto (3) NLP Group, ETIC, Pompeu Fabra University (4) ICREA.

CONCLUSION

We provided a hate speech hierarchical labeling schema that integrates the complexity of hate speech subtypes and their intersections. This allowed us to find out that distinct types of hate speech present different agreement levels between annotators. Therefore, future guidelines for annotation may benefit from specifying the particularities of the different subtypes of hate speech. Finally, in future explorations of this dataset, we will experiment with multilabel classification of hate speech to identify not only whether a message contains hate, but also the targeted groups.

Acknowledgements

This work was partially funded by the Google DNI project Stop PropagHate. Soler-Company and Wanner have been supported by the European Commission under the contract numbers H2020--7000024-RIA and H2020-786731-RIA. We would like to thank the anonymous reviewers for their insightful comments and to the annotators for their contribution to this work.

References

[1] 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.
[2] Patricia Hill Collins. 2015. Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 41:1– 20.
[3] Hatebase. 2019. Hatebase. Available in https: //www.hatebase.org/, accessed last time in February 2019.
[4] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
[5] Joni Salminen, Hind Almerekhi, Milica Milenkovic´, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning mod- els for identifying and classifying hate in online news media. In T*welfth International AAAI Conference on Web and Social Media*.
[6] Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall's. 2012. Package Various coefficients of interrater reliability and agreement.
[7] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.
[8] 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.

The author's institutes are (1) INESC TEC (2) FEUP, University of Porto (3) NLP Group, ETIC, Pompeu Fabra University (4) ICREA.

Stop PropagHate
Google DIGITAL NEWS INITIATIVE
INESCTEC
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
upf. taln Automatic Natural Language Processing Research Group