Humour identification in implicit misogyny

Iole Falsone, Marianna Guarnieri
Dept. of Interpreting and Translation
University of Bologna, Forlì
iole.falsone@studio.unibo.it
marianna.guarnieri@studio.unibo.it
September 12, 2023

Abstract

In this paper we present the workflow to create two sentence-detection models for the English language, one focused on recognizing implicit hate speech and one focused on recognizing humour. These models will be applied to a misogyny-annotated corpus in order to retrieve information concerning the presence of misogyny, implicit misogyny and misogyny disguised as humour in tweets. Developing models that can detect misogynous instances on social platforms is an issue worth concentrating efforts on in NLP, since the amount of hateful content is very high and when it is implicit it is harder to spot. Such models show interesting insights to take into account when analysing misogyny on the web, however, there is still room for improvement.

1. Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP) and its applications, the topic of hate speech detection is gaining importance. The increase of hate speech on social media explains why researchers have investigated methods to automate the detection of online hate speech.

In the present paper we will refer to "hate speech" as an umbrella term to imply abusive language that disparages or derides a particular group or individual on the basis of their religion, ethnicity, nationality, colour or gender (Gitari et al., 2015). Previous work focused mostly on explicit hate speech (Del Vigna et al., 2017; Djuric et al., 2015; Gitari et al., 2015), while recent focus has been on its implicit equivalent (El Sherief et al., 2021). This type of hate speech may not openly use offensive or discriminatory words, but it still carries harmful implications and reinforces negative stereotypes or biases relying on linguistic nuances, humour, euphemisms, metaphors (*ivi*). Detecting an explicit hate word or expression within a text is relatively easy. However, recognizing concealed hate speech within intricate nuances or contexts becomes difficult in cases where there are insufficient lexical cues (Youngwook et al., 2022).

According to Waseem and Hovy (2016), misogyny is a subset of hate speech, targeting the victim based on gender or sexuality (Samghabadi et al., 2020). Similarly, implicit misogyny can be considered as a subcategory of implicit hate speech. Online, it shows through subtle language that strengthens conventional gender stereotypes, minimises women's abilities or their contributions, playing a role in upholding gender disparities and prolonging detrimental gender norms. These subtle allusions might arise unintentionally and frequently originate

from deeply rooted cultural or societal convictions, rendering them harder to recognize and tackle (Glick et al., 1996). Detecting misogyny is a crucial NLP application because, according to the recent Online Harassment report from Pew Research Center[1], it is more likely for women to become targets of online harassment because of gender than for men (11% vs. 5%) (Pamungkas, et al., 2020).

For this reason, we will focus on detecting implicit misogyny disguised as humour, as it is one of the typical features usually seen in implicit statements of misogyny (Strathern & Pfeffer, 2022). Being misogyny a form of hate speech, we will first train a model to detect implicit and explicit hate speech, then one to detect humour. Both models will be applied to the AMI (Automatic Misogyny Identification) 2018 dataset, in order to verify if it contains examples of implicit misogyny and humour.

2. Related work

2.1 Hate speech

Several hate speech detection models have been proposed by researchers to detect hate across social platforms, addressing the task with different approaches.

Djuric et al. (2015), for instance, developed a neural language model that uses paragraph2vec to create embeddings for words and comments, trains a binary classifier using these embeddings, and then utilises the trained classifier to detect hate speech in new comments by generating embeddings for the new comments and making predictions based on those embeddings. The advantage of this approach is that it captures both the semantic meaning of individual words and the overall meaning of entire comments.

On the other hand, Gitari et al. (2015), presented a lexicon-based approach which involves identifying subjective sentences that likely contain emotional and opinionated content, building a lexicon of hate-related words using rules based on subjective and semantic features and creating a classifier that uses features from the hate-related lexicon to detect hate speech in text. The aim was to improve hate speech detection by leveraging the emotional and opinionated nature of hate speech, as well as by creating a specialised lexicon of words associated with such content. The classifier then uses this lexicon to make predictions about the presence of hate speech in given text documents.

In 2017, Del Vigna et al. tested two hate speech classifiers, one based on Support Vector Machines (SVM) and the other on Long Short Term Memory (LSTM), a Recurrent Neural Network. trying both a binary classification experiment (Hate, No hate) and a three-category classification experiment (Strong hate, Weak hate and No hate). For the binary classification, they achieved similar results to those seen in well-studied sentiment analysis tasks conducted for Italian.

As for implicit hate speech detection, El Sherief et al. (2021) created a large-scale benchmark corpus, which is a collection of text messages with detailed labels indicating the fine-grained implications of each message. This corpus served as a valuable resource for studying and understanding implicit hate speech on a significant scale. The primary purpose of this effort was to enhance the Natural Language Processing (NLP) community's understanding of and

[1] https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

ability to model implicit hate speech. They presented several advanced models as baseline approaches for identifying and explaining implicit hate speech. These neural models effectively categorised hate speech and provided more nuanced interpretations of implicit hate speech, thus giving a deeper understanding of the hateful messages.

2.2 Humour

The field of automatic humour detection is considered a complex area of research in NLP. This complexity arises due to the diverse forms of humour like irony, wordplay, metaphor, and sarcasm, coupled with the scarcity of well-defined categorizations for humour traits (Mao & Liu, 2019). Consequently, automating the recognition of humour becomes a challenging task. While some prior efforts have been made, this field still lacks a comprehensive framework for humour characterization enabling automated recognition and generation (*ivi*). As for hate speech detections, various methods can be applied to the humour detection task. One of the best performing methods is the pre-trained BERT sentence embedding model (Annamoradnejad & Zoghi, 2020; Mao & Liu, 2019; Weller & Seppi, 2019).

2.3 Misogyny

Numerous studies have been presented to identify instances of misogyny and hate speech targeted to women in the user-generated content on the web, as the proliferation of hatred towards women on social media is on the rise. Magnossão de Paula et al. (2021), in the context of the sEXism Identification in Social neTworks shared 2021 (EXIST 2021)[2] task, proposed by the Iberian Languages Evaluation Forum (IberLEF), presented a system for sexism detection and classification in English and Spanish. They compared the results of monolingual and multilingual BERT models, for identifying and classifying sexism in texts in multiple languages, experimenting with ensemble strategies, such as combining the predictions of multiple models to achieve better results. The model generated positive results, obtaining first place in both identification and classification tasks.

Evalita (Evaluation of NLP and Speech Tools for Italian) promoted in 2018 and 2020 two others shared tasks for Automatic Misogyny Identification (AMI)[3] with the purpose to automatically detect misogynistic content on Twitter, encompassing both Italian and English languages. Pamungkas et al. (2020) took part in the shared task proposing an advanced model for identifying misogyny in social media. They investigated the most predictive features to distinguish misogynistic content from not-misogynistic content and conducted a series of cross-domain classification experiments to explore how misogyny interrelates with other forms of abusive content. Finally, they carried out cross-lingual classification experiments in order to learn and generalise knowledge about misogynistic content over datasets in different languages. The results of the system proposed outperformed all state of the art systems in all benchmark AMI datasets.

3. Datasets

3.1 Implicit Hate corpus

To train the first model, we used the implicit hate corpus[4] provided by El Sherief et al. (2021). It is a collection of 22 056 tweets from the largest extremist groups in the United

[2] http://nlp.uned.es/exist2021/
[3] https://amievalita2020.github.io/
[4] https://github.com/SALT-NLP/implicit-hate

States, according to the SPLC report (2019)[5], namely Black Separatists (27.1%), White Nationalists (16.4%), NeoNazis (6.2%), Anti-Muslims (8.9%), Racist Skinheads (5.1%), Ku Klux Klan (5.0%), Anti-LGBT (7.4%), and Anti-Immigrants (2.12%). Of these, 6346 contain implicit hate speech. Tweets have been filtered removing those that were likely to be explicitly hateful, as contained explicit keywords found in NoSwear (Jones, 2020) or Hatebase (Hatebase, 2020). The folder of the corpus includes three files containing the tweets, annotated with a different criterion, each of them providing an alternative view of the data. We selected the one annotated according to three categories: explicit hate, implicit hate, not hate (implicit_hate_v1_stg1.tsv).

3.2 Humour detection

The second model was trained from the Kaggle dataset[6] for humour detection. This repository is divided into six datasets, namely: Humorous Jokes, Oneliners, Longer jokes, Reuters Headlines, English Proverbs and Wikipedia Sentences. The Python code used in the process of gathering the datasets is also made available. The content is labelled according to a binary annotation: funny, not funny. It has been divided into a training dataset (X_train, y_train) consisting of 80% of all samples from each of the above categories, and the test dataset (X_test, y_test) consisting of the remaining 20%.

3.3 AMI 2018 dataset

As previously mentioned, both models were applied to the AMI 2018 dataset[7], created for a shared task proposed by Evalita. It includes two balanced sets of tweets for both the Italian and English languages, manually annotated according to three levels: Misogyny (Misogyny vs. Not Misogyny); Misogynistic Category (Discredit, Derailing, Dominance, Sexual Harassment & Threats of Violence, Stereotype & Objectification); Target (Active vs Passive). Each corpus is divided into training (4000 tweets) and testing set (1000 tweets).
Misogynistic content from twitter has been gathered searching for manually defined representative keywords, e.g. bi**h, w**re, c*nt for English and pu****a, tr**a, f**a di legno for Italian, as well as monitoring potential victims' and identified misogynists' profiles (Fersini et al., 2018).

4. Methodology

The project was developed in Python 3[8] and is available in Jupyter format, which can be executed on the Google Colab platform. The main library used for the project is Hugging Face Transformers. We decided to fine tune the DistilBERT pre-trained model, available with Hugging Face as, according to Sanh et al. (2019) it is a “smaller, faster, cheaper and lighter version of BERT”. It is a distilled version of the original BERT model which makes it suitable for tasks where computational resources are limited. Despite being smaller, DistilBERT often retains a significant portion of BERT's performance on various NLP tasks, including sentence classification, since it is pre-trained on a large corpus, which provides it with a strong understanding of contextual language representations. It can be fine-tuned on specific tasks with a relatively small amount of specific data. It's a strong choice when a good

[5] SPLC. 2019. Hate map.
[6] https://www.kaggle.com/competitions/humor-detection/data
[7] https://amievalita2018.wordpress.com/data/
[8] https://www.python.org/download/releases/3.0/

balance between model size and accuracy is needed. Moreover, the reduced model size leads to lower memory usage, enabling it to be deployed on devices with limited memory (*ivi*).

4.1 Implicit hate detection model

4.1.1 Three-categories classification[9]

First, we uploaded the "implicit_hate_v1_stg1_posts.tsv" file from the implicit hate corpus, because it is the one annotated according to our needs. It is a table with two columns:

- Post: containing the tweets text
- Class: explicit hate, implicit hate, not hate

We excluded the other files in the corpus as they were annotated with more specific categories that we did not need for the purpose of this model.

We preprocessed the text applying lowercasing, stripping, removing hashtags, special characters, retweets and unnecessary spaces between words with the following libraries: pandas, os, regex. We enclosed all these actions in a single function. Then, we renamed the columns of the file (post, class) with the labels used by Hugging Face (text, label) and converted the annotation into numbers: explicit hate - 0 , implicit hate - 1, not hate - 2. To divide the data into training set (80%), validation set (10%) and test set (10%), we used the sklearn function `train_test_split`. With pandas, we transformed the dataframes in datasets, in the format requested by Hugging Face, and created the tokenizer with Transformers. Finally, we started the training trying different values regarding the learning rate, the batch size and the number of training epochs.

4.1.2 Binary classification[10]

For the implicit hate detection model, we also tried to train a model on a binary classification, repeating all the operations on the "implicit_hate_v1_stg1_posts.tsv" file as in 4.1.1, but this time removing the "not hate" category, to verify if we could obtain a more precise model, despite the smaller amount of training data. Being the purpose of the project to detect implicit misogyny, which falls into hate speech, we believed that the "not hate" category could be excluded.

4.2 Humour detection model[11]

In this step, we uploaded the training dataset and the test dataset from the Kaggle humour dataset. Before preprocessing, we created two pandas dataframes, one containing the train set and the other containing the test set, each set divided into the text part (x) and its corresponding label (y). The text cleaning function included almost all the actions of the previous model, except for removing hashtags and retweets, since in this case we were not dealing with tweets. There was no need to rename the annotation as it was marked with numbers: not funny - 0, funny - 1. The last operations were identical to those of the previous model: conversion of the dataframes into datasets, creation of the tokenizer, model training.

4.3 Application of the models to the AMI 2018 dataset[12]

---

[9] https://colab.research.google.com/drive/1JKPe3_dBIM0slAlNJD9vLX9FXXvfNg1o

[10] https://colab.research.google.com/drive/1dVVzQ2NqSK2suug5vuzXqWA9CChQZQHo#scrollTo=gGecKt9F6_pZ

[11] https://colab.research.google.com/drive/13GXmdFHAvd37D4uyFKUiL_EF40ylXwOi

[12] https://colab.research.google.com/drive/1IaXEkRehrSPVJzVVfbdO1tGDleWa5BHB
https://colab.research.google.com/drive/1i-uXO4cgN5yWnp8GbWxi_rekZnDgu4MR

In order to apply both models to the AMI 2018 dataset, we first uploaded the files containing the tweets in English: "en_training_anon.tsv" and "en_testing_labeled_anon.tsv". We created two dataframes with the files and merged them into one with pandas `concat` function. The resulting corpus was preprocessed with the same operations as in 4.1. Then, we imported the previous models with the best performing settings and applied them to this dataset. Finally, we added the two models classification as "hate" and "humour" columns in the dataframe and exported it as tsv file. We applied both the three-categories classification and the binary classification implicit hate speech model.

4.4 Humour identification in implicit misogyny[13]

In the last part of the project, we uploaded the tsv file with the previously classified dataset and converted it into a dataframe. We created a series of dataframes containing different combinations of classification:

- misogynous_df, containing only the content labelled as misogynous
- implicit_misogyny_df, from the misogynous dataframe containing the implicit hate content only
- humour_in_misogyny_df, from the misogynous dataframe with the funny content only
- humour_in_implicit_misogyny_df, from the implicit_misogyny_df containing also the funny content

This could have been done with other methods, like the `count` function in pandas, but we decided to create different dataframes so that it would be easier to download the filtered data.

5. Results

In the process of training the models, we experimented different combinations of batch sizes and learning epochs, noticing that:

- for the hate models, as presented in Table 1, the best result retrieved with the three-categories classification was an accuracy of 0.738 obtained with a batch size of 128, 20 epochs and a learning rate of 3e-6, while the binary classification model reached an accuracy of 0.866, with the same batch size, 15 epochs and a learning rate of 3e-7. El Sherief et al. (2021) with BERT reached an accuracy of 0.783 with the binary classification, however, we can't state with confidence that our model outperformed theirs, as it does not always recognise implicit hate, probably due to the small amount of data used in the training.

Table 1. Implicit hate models results

Model	**Training Loss**	**Validation Loss**	**Accuracy**
Three-categories	0.538	0.646	0.738
Binary	0.597	0.429	0.866

[13] https://colab.research.google.com/drive/1UMxn5T8NB5nhvKpHUdI1Zp0YWno0h5fb
https://colab.research.google.com/drive/16YN7oqXed_Qgxry2tpHIthDHI6w-RqGb

- for the humour model, as shown in Table 2, the best result retrieved was an accuracy of 0.972 obtained with a batch size of 128, 20 epochs and a learning rate of 3e-6. This can be considered an acceptable result, being 0.982 the benchmark of the accuracy in humour detection (ColBERT model).

Table 2. Humour model training results

Model	**Training Loss**	**Validation Loss**	**Accuracy**
Humor	0.072	0.090	0.972

The models application to the AMI dataset took about 15 minutes.
Being the total number of instances in the AMI 2018 Dataset 5000, of which 2245 (44.9%) classified as misogynous, the results of the classification are the following:

Table 3. AMI 2018 Dataset classification

Implicit hate model used	**Implicit misogyny**	**Humour**	**Humour in implicit misogyny**
Three-categories classification	44.54% (1000)	95% (2134)	43.43% (975)
Binary classification	100% (2245)	95% (2134)	95% (2134)

The classification seems to suggest that almost all of the content of the AMI dataset is humorous, however, this information is not fully reliable as the humour model is not trained on tweets, so, it is possible that it is not able to recognise humour in tweets with confidence.
For the three-categories hate model, we cannot say for sure that all data have been classified correctly as, even in the model training step, we could not always obtain the right classification when the model was tested with a sample sentence. The percentages of the binary model demonstrate that the amount of training data was too small, so, even if the accuracy was quite high, it is clear that the model is not able to distinguish external instances.

6. Conclusion

The present study highlights the challenges in developing accurate and reliable models to detect instances of humour intertwined with implicit misogyny and, also, how essential it is to improve such methods, in order to identify the perpetuating of harmful stereotypes under the guise of humour in social media.
The models we trained are not completely reliable and accurate in labelling the sentences the way they are intended to, however, we believe that they are a good starting point and that could be improved, including more training data and trying further combinations of parameters.

References

Annamoradnejad, I., & Zoghi, G. (2020). Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*, *1*(3).

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., ... & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)* (pp. 86-95).

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web* (pp. 29-30).

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR workshop proceedings* (Vol. 2263, pp. 1-9). CEUR-WS.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, 70(3), 491-512.
Hatebase. 2020. [link].

Jones, R. (2020). List of swear words, bad words, curse words - starting with a.

Magnossão de Paula, A. F., da Silva, R. F., Schlicht, I. B. (2021). Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models. Volume 2943. 356-373. *arXiv preprint arXiv:2111.04551*.

Mao, J., & Liu, W. (2019, September). A BERT-based Approach for Automatic Humor Detection and Scoring. In *IberLEF@ SEPLN* (pp. 197-202).

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study, *Information Processing & Management*, 57 (6)

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Samghabadi, N. F., Patwa, P., PYKL, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France. European Language Resources Association (ELRA).126–131.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*. /abs/1910.01108

Strathern, W., & Pfeffer, J. (2022). Identifying Different Layers of Online Misogyny. *arXiv preprint arXiv:2212.00480*

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Weller, O., & Seppi, K. (2019). Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. (2022). Generalizable Implicit Hate Speech Detection Using Contrastive Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.