Violentometer: measuring violence on the Web in real time

Henrique S. Xavier
hxavier@nic.br
NIC.br
São Paulo, SP, Brazil

Figure 1: An expressive oil painting of a giant thermometer stuck in the middle of a bunch of angry monsters fighting each other, according to the generative AI DALL.E.

ABSTRACT

This paper describes a system for monitoring in real time the level of violence on Web platforms through the use of an artificial intelligence model to classify textual data according to their content. The system was successfully implemented and tested during the electoral campaign period of the Brazilian 2022 elections by using it to monitor the attacks directed to thousands of candidates on Twitter. We show that, despite an accurate and absolute quantification of violence is not feasible, the system yields differential measures of violence levels that can be useful for understanding human behavior online.

CCS CONCEPTS

• **Applied computing** → *Sociology*; Document analysis; • **Information systems** → Social networks; • **Computing methodologies** → *Information extraction.*

KEYWORDS

violence, hate speech, content analysis, computational social science, automation, neural networks

ACM Reference Format:
Henrique S. Xavier. 2023. Violentometer: measuring violence on the Web in real time. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587364

1 INTRODUCTION

The existence of violence and hate speech on web platforms is known [16], as well as its more specific form of electoral violence, that is, violence directed to electoral candidates. Several studies indicate that electoral violence in Brazil has grown over time, both in offline and online spaces [2, 7, 10]. These types of manifestation affect all candidates, but are especially harmful when directed at marginalized social groups by serving as yet another instrument for strengthening their marginalization [1].

At the same time that the Web intensifies communication flows and textual data production – as well as the dissemination of political violence –, its machine-readable environment makes it possible to automate the monitoring of content. The combination of these two characteristics makes the Web a favorable and promising terrain for the application of Natural Language Processing (NLP) and Artificial Intelligence (AI) on statistical studies regarding aggression in the political landscape. Such studies can reveal new dynamics about politics and violence on and off the Web, raise awareness on important social issues and yield methods for detecting relevant events in real time.

This work describes the Violentometer, an AI application for quantifying in real time the amount of electoral violence on Twitter first used during the Brazilian 2022 elections to raise awareness of political violence against women. The system ran continuously during the electoral campaign period, from August 16th to October 30th, measuring every three hours the level of violence on tweets directed to the candidates. A daily summary of the results was automatically published on a webpage at https://vamosjuntasorg.github.io/violentometro.

The Violentometer project was released online on the public repository https://github.com/cewebbr/violentometro, under the GNU General Public License (GPLv3), following open science best practices. It will also go live during the Web Conference 2023, for demonstrating purposes.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587364

This paper is organized as follows: Section 2 starts with an overview of the Violentometer system, followed by a description of the AI model employed to identify violence in texts and its training process (Section 2.1). Section 2.2 explains the code that monitors Twitter and captures the posts, along with its strategy for dealing with a large amount of target users; and Section 2.3 explains how the model's predictions are turned into a measurement of the amount of violence on the Web platform at a given time. A test on the capacity of the system to measure violence is presented in Section 3, along with a discussion on its limitations. Finally, Section 4 presents some use cases and results obtained with the system and our final considerations.

2 SYSTEM DESCRIPTION

The Violentometer system is a scheduler that runs, every few hours, a pipeline for measuring the level of violence in a given Web platform (in this particular implementation, Twitter). The pipeline captures publications made on the platform, associated to a provided list of users, applies an AI model to compute the probability that each text is violent and, given these probabilities, estimate the amount of violent texts posted in a given time window. The most important parts of the system are described in the subsections below. The results may be displayed automatically online by updating a webpage (as was done at https://vamosjuntasorg.github.io/violentometro).

Disruptions to the dataflow may occur in basically two ways: limitations on the access to an online resource (i.e. the Twitter API) or an error caused by ill-formed collected data. The system's architecture is designed to be robust against such failures – so that an eventual error does not take the whole system down – essentially by partitioning or isolating the process into independent parts. More details are given below.

2.1 Violence detection model

The AI model used in this work to estimate the degree of violence in tweets was built from the pre-trained BERTimbau model [15]. This model uses the BERT architecture – Bidirectional Encoder Representations from Transformers [4] – and was pre-trained through the self-supervised tasks of Next Sentence Prediction and Masked-Language Modeling using 3.53 million documents in Portuguese, containing 2.68 billion tokens, obtained from the Brazilian Web (the brWaC corpus) [17].

The pre-trained model was fine tuned for the binary classification of texts as violent or non-violent using an annotated set of texts in Portuguese composed of: 1,250 comments made by internet users on the Brazilian news website http://g1.globo.com, annotated as offensive or not [3]; and 5,668 tweets containing keywords related to hate speech, collected between January and March 2017, annotated as hate speech or not [5].

Table 1 compares the BERTimbau model described above with a bag-of-words machine learning (ML) model and with a random classification in terms of several metrics computed over a test set (i.e. a randomly selected subset of the data that was not used for training nor for fine-tuning the model). It shows that the BERTimbau model outperforms the best-fit bag-of-words model, thus being the best option for identifying violent texts among the ones tested.

Table 1: Comparison of classification metrics between the BERTimbau model, a bag-of-words ML model and a random classification.

Metric	BERTimbau	Bag-of-words	Random
Accuracy	0.901	0.872	0.510
F1	0.596	0.471	0.281
Precision	0.689	0.736	0.185
Recall	0.525	0.346	0.584

2.2 Twitter monitoring system

The Twitter monitoring system is a code in Python that uses the "mentions" endpoint of the Twitter API[1] to capture, every 3 hours, the tweets produced in the last 3 hours in response to or that mention a set of Twitter profiles belonging to electoral candidates. The API call, made once per candidate and instant in time, is isolated in a Python try except block so that eventual call errors do not affect the whole system.

It is important to note that the endpoint may not return some tweets that, in principle, it should.[2] In addition, the endpoint response is limited to the 800 most recent tweets, which makes the capture incomplete when more than 800 mentions are made in a period of 3 hours. To minimize this issue, the capturing period could be decreased. In any case, the effects of an eventual incompleteness are considered when counting tweets, as explained in Section 2.3.

Given that the goal of the system is to quantify the amount of violence directed to candidates in general, that the number of candidates, for all positions (including representatives at all government levels), is quite large (a few thousands), and that the Twitter API has a consumption cap, a statistical approach was used: every 3 hours, the system draws a third of the candidates' Twitter profiles and carry out the capture as described above. To avoid duplication of captured tweets, the profiles used in the previous capture are ignored when drawing the profiles for the next capture. This strategy introduces a dependence between successive captures that can be erased by aggregating the results over a larger time period.

All raw data collected with the Twitter API – and the capture logs – are saved locally. The code that applies the AI model to the texts is ran as a separate process, thus preventing data reading errors from halting the capture process. This independence also allows for both tasks to be performed asynchronously and for the possibility of updating the classifications made. The aggregation of data for the quantification of the violence is also performed by an independent process, following the same reasoning.

2.3 Quantifying violence

The goal of Violentometer is to quantify violence on a Web platform by counting the number of violent texts published on that platform in a given period of time. Since the actual counting is impossible due to the content loss mentioned in Section 2.2, the prediction

[1] https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-mentions

[2] https://twittercommunity.com/t/mentions-timeline-json-endpoint-may-miss-some-tweets/178852

errors of the model and the subjective nature of the concept of violence, the following estimator is used instead:

$$\hat{n}_t = \frac{|C|}{|c_t|} \sum_{k \in c_t} \frac{\Delta T_0}{\Delta T_{kt}} \sum_{i \in s_{kt}} p_i. \tag{1}$$

On the equation above: $\hat{n}_t$ is the estimated number of violent tweets directed to candidates in set C at a given time t; c_t is the subset of randomly sampled candidates from C at time t whose captures of mentions were successful; ΔT_0 is the capture time window (in this case, 3 hours); ΔT_{kt} is the time period actually covered by the API response when capturing tweets at time t mentioning the candidate k, given that the response is limited to the 800 most recent tweets; s_{kt} is this set of captured tweets; and p_i is the probability given by the model that the tweet i is violent. $\Delta T_{kt} = \Delta T_0$ unless the endpoint limit is reached, in which case ΔT_{kt} is smaller than ΔT_0 and given by the time interval between the first and last tweets in s_{kt}.

Eq. 1 assumes that c_t is statistically representative of C; the same assumption is used for the tweets captured in the time interval ΔT_{kt} with respect to those that would be captured if the endpoint response was not limited to the 800 most recent tweets. Lastly, we have empirically verified that the sum of p_i is a better estimator of the number of violent texts in s_{kt} than the counting of tweets predicted to be violent – i.e. it is more accurate and precise (it has a smaller variance and bias).

The subjectivity of the flagging of texts as violent is partially resolved through aggregation and the use of p_i. By construction, texts that are controversial to annotate tend to get assigned intermediary probabilities, and a sample of many such texts will contain an intermediary fraction of texts annotated as violent even if classifications made by different people to individual instances vary. However, as Section 3 explains, some subjectivity remains.

2.4 Identifying the target of the comment

During the implementation of the project for the Brazilian 2022 elections, is was realized that approximately half of the violence detected in replies to candidates are not directed to the candidates themselves but to adversaries mentioned (and often criticized) by them. The AI model described in Section 2.1, however, does not differentiate these two situations.

Given that the goal of the implementation was to quantify the aggression done directly to the candidates, an auxiliary ML model was built to identify if the candidate being replied was the object of the comment or not. The product of the probability computed by this targeting model with the probability of violence computed by the AI model from Section 2.1 replaced p_i in Eq. 1. In future projects, depending on its goal, either this or the original approach may be adopted.

3 SYSTEM VALIDATION

In order to check that the AI model was able to estimate the number of violent texts in a sample, 1.094 captured tweets were manually annotated by at least 3 specialists, and their final classes were set by majority voting. These tweets – completely new instances with no relation to the dataset used to train, validate or test the model – were binned according to the probabilities given by the model, and the fraction of tweets manually classified as violent in each bin was computed. Figure 2 shows the relation between the model's predictions and the human assessment, where three characteristics can be noted: a linear relation is an excellent fit to the data; the relation is monotonic, i.e. an increase in the predicted probability always translates into an increase in the amount of violent texts; the relation is biased, that is, the exact amount of violent texts is, in general, slightly different than the predicted amount.

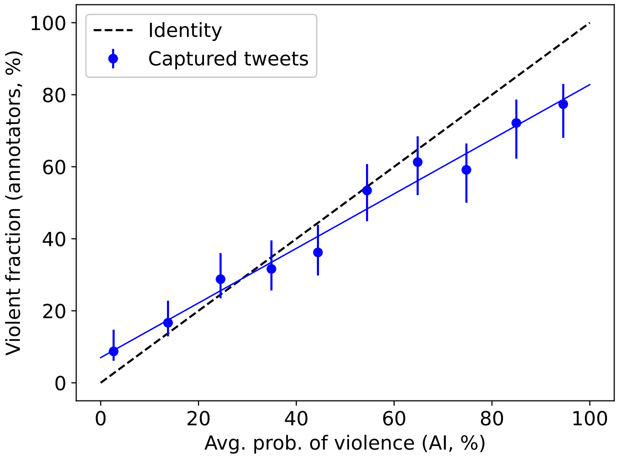

Figure 2: Relation between the average probability that a tweet is violent, according to the AI model (horizontal axis), and the fraction of tweets that are considered violent by the majority of a group of specialists (vertical axis). The data points represent 1.094 annotated tweets and the solid blue line a best linear fit to the data. The dashed black line represents the identity relation. The error bars represent 90% confidence intervals computed under the assumption that the data follow binomial distributions.

These characteristics mean that, despite not providing an exact absolute scale for the number of violent texts posted at a given time on a Web platform, the system is able to detect increases and decreases in the amount of violent texts through time. The system works in a similar fashion to a thermometer, which measures the average kinetic energy of particles on a scale with arbitrary zero point and scale factor. On top of that, the bias between predicted and true violence levels is small enough that a rough absolute estimate can be made. For instance, Figure 2 shows that when the average predicted violence reaches 95%, the fraction of texts that are violent reaches 80%; while a prediction of 5% correponds to a true fraction of 11%.

At least part of the observed bias is caused by subjectivity in the definition of violence. We verified that changes in the pool of specialists do change the intercept and slope of the relation in Figure 2. However, the relation is always monotonic, meaning that different people might not agree on the measured absolute level of violence but they will agree with respect to differential changes in this level.

4 RESULTS AND FINAL REMARKS

Violentometer is a online violence monitoring system that was applied to the context of the Brazilian 2022 elections. Contrary to previous investigations of violence dynamics during (US) elections that used word embeddings [9], dictionary methods and bag-of-word models [13], it employed the BERT transformer – an approach that was well evaluated in previous works [8, 14]. It highlighted that 975 female candidates in the Brazilian 2022 elections were, together, attacked about 2.000 times every day, and that these attacks are highly concentrated on the most prominent candidates.

The system helped uncovering important violence dynamics. For instance, it showed that violence on social media may come as bursts in response to a candidate's comments about other candidates, specially criticisms against adversaries; a similar pattern was observed in the context of US elections [13]. Depending on the original comment, two reactions may occur: commentator's supporters may join in the critique in a violent manner, cursing the adversary; or the adversary's supporters may attack the commentator in response to its comment.

The comments and responses can (and often) reflect offline events, as noticed during US elections [13]. The Brazilian 2022 elections was a highly polarized process, specially for the presidential run. Violentometer showed that major bursts of violence against candidates on Twitter were associated to dates and times of presidential debates and interviews on TV, as well as the actual day of the election. The mechanism relating the online reactions to offline events involved comments made by other candidates with respect to the presidential ones, in the manner described in the previous paragraph. Such relations can be used to detect relevant events and to study the interactions between the online and offline worlds.

There are several papers that highlight the subjective nature of violence and hate speech annotation, and that distinct social groups may evaluate sentences differently with regard to this characteristic [11, 12]; an aspect also observed in our study (see Section 3). However, we argue that Violentometer provides reliable differential (i.e. relative) measurements of violence levels since the relation between model and human assessments is monotonic and linear. It is important to state that this assumes the AI model to be similarly well adjusted to the situations being compared. Comparing situations in which the vocabularies and grammar structures are different – such as texts produced by distinct social groups – or when the targeted social groups are different may produce biased results [6, 11]. The implementation made during the campaign period of the Brazilian 2022 elections, when comparisons were made between two different times within an interval of a few months for a fixed set of targeted profiles, is an example of a use case in which the model is expected to provide unbiased relative results.

In the future, important contributions to this work may come in the form of: improving the model's performance through the use of data augmentation and other data processing techniques; and further investigating the limitations on violence level measurements caused by annotator replacement and changes in targeted social groups. The latter goal can be tackled through an increase in the amount of annotated data and the use of a variety of analysis strategies, as a way to ensure robustness in the results.

REFERENCES

[1] Revista AzMina and Internetlab. 2021. *MonitorA: relatório sobre violência política online em páginas e perfis de candidatas(os) nas eleições municipais de 2020.* https://internetlab.org.br/wp-content/uploads/2021/03/5P_Relatorio_MonitorA-PT.pdf

[2] Felipe Borba. 2022. Julho–Setembro 2022. *Boletim Trimestral do Observatório da violência política e eleitoral no Brasil* 11 (2022). http://giel.uniriotec.br/files/BoletimTrimestralnÂž11-Julho-Agosto-Setembro2022.pdf

[3] Rogers P. de Pelle and Viviane P. Moreira. 2017. Offensive Comments in the Brazilian Web: a dataset and baseline results. (2017).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3).*

[6] Alessandra Gomes, Dennys Antonialli, and Thiago Oliva. 2019. *Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet?* https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/

[7] Alice Maciel, Anna Beatriz Anjos, Caroline Farah, Ciro Barros, Ethel Rudnitzki, José Cícero da Silva, Julia Dolce, Laura Scofield, Mariama Correia, Rafael Oliveira, Raphaela Ribeiro, Rute Pina, Gabriella Soares, Rogerio Galindo, Inara Fonseca, Paula Guimarães, Vitória Régia, Débora Britto, Paulo Eduardo Dias, Liana Melo, Mirella Lopes, and Rafael Duarte. 2020. Eleições municipais provocaram cinco casos de violência política por dia em novembro. *Agência Pública* (2020). https://apublica.org/2020/12/eleicoes-municipais-provocaram-cinco-casos-de-violencia-politica-por-dia-em-novembro/

[8] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *Complex Networks and Their Applications VIII*, Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha (Eds.). Springer International Publishing, Cham, 928–940.

[9] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. 2017. Measuring Offensive Speech in Online Political Discourse. In *7th USENIX Workshop on Free and Open Communications on the Internet (FOCI 17).* USENIX Association, Vancouver, BC. https://www.usenix.org/conference/foci17/workshop-program/presentation/nithyanand

[10] Alexsandro Ribeiro, Carolina Zanatta, Caroline Farah, Gabriele Roza, José Lázaro Jr., Mariana Simões, and Thays Lavor. 2018. Violência eleitoral recrudesceu no segundo turno. *Agência Pública* (2018). https://apublica.org/2018/11/violencia-eleitoral-recrudesceu-no-segundo-turno/

[11] Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. 2022. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22).* Association for Computing Machinery, New York, NY, USA, 1585–1603. https://doi.org/10.1145/3531146.3533216

[12] Yisi Sang and Jeffrey Stanton. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Springer International Publishing, Cham, 425–444.

[13] Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. Trumping hate on Twitter? Online hate speech in the 2016 U.S. Election campaign and its aftermath. *Quarterly Journal of Political Science* 16, 1 (2021), 71–104. https://doi.org/10.1561/100.00019045

[14] Félix Silva and Larissa Freitas. 2022. Brazilian Portuguese Hate Speech Classification using BERTimbau. *The International FLAIRS Conference Proceedings* 35 (May 2022). https://doi.org/10.32473/flairs.v35i.130594

[15] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer International Publishing, Cham, 403–417.

[16] Alexander T. Vazsonyi, Daniel J. Flannery, and Matt DeLisi. 2018. *The Cambridge Handbook of Violent Behavior and Aggression* (2 ed.). Cambridge University Press. https://doi.org/10.1017/9781316847992

[17] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1686