Conflict Prediction in Public Administration: A Gender-Aware Evaluation Framework

Sylvie Cerise

Università della Valle d'Aosta - Université de la Vallée d'Aoste, Aosta, Italy

Abstract

Understanding and predicting conflict within institutional settings is essential for promoting more equitable and effective decision-making processes. This paper introduces a novel approach to conflict prediction by incorporating gender dynamics and applying the SAFE AI framework - a multi-dimensional evaluation tool that assesses machine learning models based on four key dimensions: accuracy, fairness, robustness, and explainability. Focusing on legislative debates within the Regional Council of Aosta Valley, we use sentiment analysis through a BERT-based model to explore how gender influences emotional tone, participation patterns, and conflict escalation. Our analysis reveals that female councilors tend to express more negative sentiments and intervene less frequently than their male counterparts. We compare multiple machine learning models—including Naive Bayes, Logistic Regression, Decision Trees, Random Forest, and XGBoost—for their ability to predict sentiment polarity. While Random Forest demonstrates strong predictive accuracy and robustness, it also achieves a balanced performance across fairness and explainability, making it a well-rounded choice under the SAFE AI framework. These findings underscore the importance of considering gender dynamics in predictive conflict models and highlight the need for ethical frameworks in evaluating machine learning systems, particularly in contexts where fairness and transparency are critical.

Keywords

SAFE AI, sentiment analysis, conflicts, public administration, gender.

1. Introduction

Understanding conflict is particularly important given its prevalence in human relationships and social structures. While armed conflicts have received substantial scholarly attention - likely due to their profound geopolitical consequences - the conflicts most commonly encountered in everyday life are non-violent in nature. These include interpersonal or institutional disputes that, despite lacking physical violence, can significantly influence social dynamics, organizational functioning, and even the stability of governance structures. Such conflicts, though often less visible and less intense than violent confrontations, can have profound long-term effects on cooperation, trust, and the efficiency of decision-making processes within organizations and institutions.

While traditional conflict prediction models have predominantly focused on identifying the sources of conflict - such as resource scarcity, misaligned goals, or communication breakdowns - they often fail to account for critical social dynamics that influence the escalation and resolution of conflicts. A major aspect of these dynamics, which is still underexplored, is the role of gender. In many conflict situations, gender disparities in communication styles, decision-making, and participation can significantly affect the trajectory of disputes. Despite its potential to shape both the occurrence and resolution of conflicts, gender is often overlooked in traditional conflict prediction models.

This paper introduces the SAFE AI framework [1], a novel approach to conflict prediction that evaluates machine learning models across four dimensions: accuracy, fairness, robustness, and explainability. Moving beyond traditional metrics like accuracy alone, SAFE AI ensures that predictive models are not only effective but also ethically sound. By including fairness and robustness, it offers a more comprehensive evaluation, helping mitigate biases - particularly important in contexts such as public administration, where institutional and gender dynamics significantly impact outcomes. We apply this

2nd Workshop "New frontiers in Big Data and Artificial Intelligence" (BDAI 2025), May 29-30, 2025, Aosta, Italy

s.cerise4@univda.it (S. Cerise)

0009-0001-7149-0188 (S. Cerise)

framework to evaluate machine learning models predicting conflict escalation in political discourse, providing a holistic, ethically grounded assessment that balances predictive power with fairness.

Our study focuses on conflict dynamics in legislative debates within the Regional Council of Aosta Valley. Using sentiment analysis of these debates, we examine how gender influences aspects of conflict, including escalation, participation styles, and emotional tone. This approach uncovers gendered communication patterns and conflict behaviors that might otherwise go unnoticed. By applying the SAFE AI framework to assess various machine learning models, we aim to identify those that are not only the most accurate but also the most fair, robust, and transparent, minimizing gender disparities and ensuring explainable, justifiable decision-making.

The paper is structured as follows: Section 2 provides the theoretical background, discussing the conflict process model and the role of gender in shaping conflict dynamics. Section 3 outlines the methodology, focusing on the SAFE AI framework and its application to conflict prediction models. Section 4 describes the dataset used in the study, including details about its scope and sources. Section 5 presents preliminary results, comparing the performance of different models and examining the gender-based differences in conflict predictions, and limitations of the study. Finally, Section 6 discusses the implications of these findings for future research in conflict prediction, with a particular focus on ethical AI and the integration of gender considerations into predictive models.

2. Background on conflict

Conflict is a fundamental aspect of human interaction, present across all organizational and social structures. Though often viewed negatively, conflict can have positive organizational outcomes—promoting better decision-making, self-awareness, and adaptability in rapidly changing environments [2, 3].

However, conflict also poses risks. It can consume time, distract employees, induce stress, and weaken team cohesion by discouraging the sharing of vital information [4]. Despite extensive study, conflict remains complex, typically involving opposing interests recognized by the involved parties, shaped by perceptions and past experiences [3, 4].

In institutional contexts like public administration, conflict may stem from political disagreements, competition for resources, or inefficiencies. The conflict process model [3] highlights two main escalation drivers: sources of conflict and conflict management styles. Miscommunication is a particularly potent and often underestimated source, while Thomas's model [5] outlines five management styles based on assertiveness and cooperativeness.

Gender adds another dimension. Inequities in representation, education, and economic access have been linked to broader societal tensions [6, 7]. Communication is also gendered; as Gray [8] notes, men and women often have distinct patterns that shape how conflict is expressed and resolved, particularly in hierarchical settings.

Given the central role of communication in conflict, this study analyzes council session discourse to identify relational tensions and communication breakdowns. It also examines how gendered communication influences conflict in formal settings.

To support this analysis, we employ a range of methodological tools used in conflict prediction, from traditional statistical models to more recent machine learning (ML) and deep learning (DL) approaches. Early work often relied on logistic regression, valued for its interpretability in identifying key predictors [7]. More advanced models, such as dynamic multinomial logit frameworks, account for temporal and contextual factors like prior conflicts or regional influences [9, 10].

In the ML domain, Random Forests (RF) have demonstrated strong performance, especially in handling complex, non-linear relationships and dealing with noise, missing data, and multicollinearity [11, 12]. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, offer promise for sequential data analysis. However, their effectiveness often depends on large training datasets, and they have not consistently outperformed RF models in large-scale conflict prediction tasks [12].

3. Methodology

3.1. BERT

To analyze the emotional tone of political discourse, we use BERT (Bidirectional Encoder Representations from Transformers) [13], a transformer-based model known for its ability to capture the context and meaning of language more effectively than traditional NLP approaches [14]. Specifically, we adopt the bert-base-multilingual-uncased-sentiment model, a pre-trained variant fine-tuned for sentiment analysis across six languages. Its architecture - 12 transformer layers, 768 hidden units, and 12 attention heads - enables it to grasp complex linguistic patterns, making it well-suited for multilingual political text. The model assigns each intervention a sentiment score on a five-point scale (from very negative to very positive), offering a nuanced view of the emotional dynamics in the debate. This allows us to trace shifts in tone, detect moments of tension or consensus, and link emotional trends to speaker profiles and institutional dynamics, with a specific focus on gender differences in communication style and conflict escalation.

3.2. SAFE AI package

To evaluate the models applied to our dataset, the SAFE AI package [1] was used. SAFE AI provides a unified framework for assessing model performance along four critical dimensions: accuracy, fairness, robustness, and explainability. These dimensions are evaluated through a consistent methodology based on Lorenz curves and the concordance curve, which respectively measure the inequality and consistency of prediction distributions. Lorenz curves are used to assess how concentrated predictions are (e.g., whether a model's outputs are disproportionately distributed across a few categories or groups), while the concordance curve captures agreement between model predictions and ground truth, or between predictions under input perturbations. This setup enables a comparative analysis of model behavior across subgroups, highlighting potential biases, overfitting, or instability. In contexts like policy-making, where model transparency is essential, SAFE AI helps balance performance with ethical accountability—ensuring that high accuracy does not come at the cost of fairness or interpretability.

4. Dataset

To study conflict dynamics in a structured and empirical way, we are building a dataset based on the official transcripts of the Regional Council of Aosta Valley. The final version will include nearly 50,000 documents spanning over four decades (1981–today). However, in this first stage, we focus on a more recent and manageable subset: more than 5,000 legislative discussions from 2018 to 2024, covering the last two legislative terms. This subset serves as a test-bed to develop and validate our feature extraction methods and analytical pipeline.

4.1. Data

The dataset draws from two primary sources that together offer a rich foundation for analyzing political conflict. The first source comprises official transcripts of council meetings, which provide complete, verbatim records of legislative debates and discussions. These documents are especially rich because they capture the actual flow of political interaction - speeches, interruptions, emotional tones, and argumentation styles - which are crucial to analyzing conflict. Such nuances are essential for understanding how conflict unfolds in real time. The second source consists of detailed socio-demographic data about council members, including variables such as gender, age, and political affiliation. This enables researchers to link discourse to individual identities, and to explore how personal characteristics and positional authority influence patterns of behavior, participation, and conflict during deliberations.

Table 1
Comparison of the sentiment analysis between gender

	mean very_negative	mean negative	mean neutral	mean positive	mean very_positive
Female	0.171	0.265	0.255	0.204	0.105
Male	0.165	0.251	0.254	0.218	0.112
p-value	6.6e-3	1.9e-17	3.5e-1	1.6e-15	4.0e-6

4.2. Features

To explore how conflicts unfold in political discussions, we extract three main types of features from the transcripts:

1. Socio-demographic features include each speaker's gender, age, seniority, and political affiliation. These help us investigate how different profiles might relate to participation style, conflict escalation, or rhetorical strategies. They are also key to assessing representational gaps and power asymmetries.
2. Structural features describe how the conversation is organized. At the macro level, we analyze the total length of a discussion, the number of interventions, and whether a formal vote took place. At the micro level, we look at individual contributions - who spoke, for how long, and in what order. This helps us map the intensity and rhythm of exchanges, and detect patterns such as dominance, persistence, or marginalization.
3. Sentiment and emotional features capture the tone of each intervention. Using a multilingual BERT-based sentiment model, we assign a sentiment score to every speech act, ranging from very negative to very positive. In this study, sentiment serves as a proxy for conflict escalation: we assume that emotionally negative interventions are more likely to reflect or contribute to escalating tensions. This enables us to detect emotionally charged moments, monitor shifts in tone, and identify phases of heightened antagonism or reconciliation within the discussion.

Together, these features form a rich and multi-layered dataset that can support both quantitative and qualitative analyses of institutional conflict, with a special focus on how gender and power shape political communication.

5. Preliminary results

The preliminary analysis on the dataset highlights noticeable gender-based differences in political communication. On average, male councilors intervened more frequently and with longer speeches compared to their female colleagues - averaging 4.02[1] interventions with a mean length of 3,390 characters, versus 2.46 interventions averaging 3,296 characters for women - pointing to potential disparities in speaking time and participation.

From a sentiment perspective, as shown in Table 1, female interventions displayed higher levels of negative and very negative sentiment, while male interventions leaned slightly more toward the positive end of the spectrum. Neutral sentiment remained roughly similar across genders. Statistical tests confirm that these differences are significant in several sentiment categories, suggesting that gender may influence both the emotional tone and structural characteristics of discourse.

To evaluate the predictive capacity of the dataset, we tested a range of classification models aimed at predicting the sentiment polarity (positive vs. negative) of each intervention. For this purpose, we categorized each instance based on the combined scores of the "positive" and "very positive" classes versus the "negative" and "very negative" ones. We then retained only the most emotionally polarized interventions - those where either the "very positive" or "very negative" score exceeded 0.2 - ensuring the focus was on interventions with a clear emotional stance. The models included Naive Bayes, Logistic

[1]The frequency is calculated on average over the male population and over the female population separately

Table 2
Comparison of the S.A.F.E. metrics across models

	RGA (accuracy)	gender RGF (imparity)	mean RGR (robustness)	min RGR (robustness)	mean RGE (explainability)	gender RGE (explainability)
Naive Bayes	0.525	0.024	**0.982**	0.826	0.045	6.1e-7
Logistic Regression	0.579	0.055	0.976	0.937	0.034	4.8e-2
Decision Tree	0.678	0.037	0.968	0.916	**0.135**	**1.2e-1**
Random Forest	0.719	**0.015**	0.980	**0.956**	0.084	8.0e-2
XGBoost	**0.722**	0.025	0.979	0.934	0.071	5.2e-2

Regression, Decision Tree, Random Forest, and XGBoost. All models were trained using an 80/20 train-test split and validated through cross-validation to ensure robustness and generalizability.

Table 2 presents model performance using the SAFE AI evaluation framework, which assesses accuracy (RGA), fairness (gender RGF), robustness (RGR), and explainability (RGE). Among all models, XGBoost achieved the highest accuracy (RGA = 0.722), closely followed by Random Forest (0.719), both outperforming simpler models like Logistic Regression (0.579) and Naive Bayes (0.525).

Interestingly, despite its high accuracy, Random Forest also showed the lowest gender disparity (gender RGF = 0.015), indicating that its predictions were more consistent across male and female councilors. Logistic Regression, while often considered fairer in theory, exhibited a higher gender disparity (0.055), although it still performed better than Naive Bayes in both fairness and accuracy.

In terms of robustness, Naive Bayes had the highest average robustness (mean RGR = 0.982), indicating overall stability under feature perturbations. However, Random Forest had the highest minimum robustness (min RGR = 0.956), meaning even its most sensitive feature had limited impact on prediction when perturbed. This makes Random Forest particularly resilient across all input features.

When it comes to explainability, the Decision Tree model stood out with the highest scores in both overall (mean RGE = 0.135) and gender-specific explainability (gender RGE = 0.12). This suggests its predictions are more traceable to individual features and that gender plays a more interpretable role in its decision-making. In contrast, Naive Bayes and Logistic Regression scored lower on explainability, despite being traditionally viewed as interpretable models, likely due to the complexity of the decision boundary in this context.

Overall, Random Forest emerged as the most balanced model—combining high accuracy, low gender bias, strong robustness, and reasonable interpretability—making it a strong candidate for deployment in sensitive institutional settings.

5.1. Limitations

While the preliminary results provide valuable insights into gendered patterns in political discourse, several limitations must be acknowledged.

First, the dataset is geographically constrained to the Aosta Valley region, which may limit the generalizability of the findings to broader national or international contexts. Local political dynamics, institutional structures, and demographic compositions could uniquely shape communication styles, and these factors may not be representative of other settings.

Second, the temporal scope of the data (2018–2024) presents both strengths and weaknesses. While it ensures consistency by avoiding major historical shifts, it also restricts the ability to capture longer-term trends or generational changes in political discourse. Extending the time frame could offer a richer longitudinal perspective but may also introduce biases related to evolving sociopolitical contexts, such as shifting norms around gender or changes in institutional rules.

Finally, sentiment analysis, while useful for large-scale text classification, is an imperfect proxy for political conflict or tension. Emotional tone does not always map neatly onto substantive disagreement or opposition, and sentiment classifiers may overlook nuance, sarcasm, or context-specific rhetoric. As such, caution is warranted when interpreting sentiment polarity as a direct indicator of political confrontation or alignment.

6. Conclusion

This study presents a novel contribution to conflict prediction in institutional settings by integrating gender dynamics with machine learning evaluation using the SAFE AI framework. Through sentiment analysis of legislative interventions and the inclusion of socio-demographic variables, our findings highlight significant gender-based differences in political discourse—particularly in speaking time, emotional tone, and participation. These insights suggest that gender plays a meaningful role in shaping how conflict manifests and is communicated within political institutions.

By employing the SAFE AI framework, we go beyond conventional performance metrics to assess models along key ethical dimensions—fairness, robustness, and explainability. This multidimensional approach is especially critical in public sector applications, where the implications of algorithmic decisions can reinforce or mitigate existing inequalities. Our results identify Random Forest as a particularly well-balanced model, offering strong predictive performance with minimal gender bias and high resilience to perturbations.

Importantly, this research should be seen as a first step in a broader and ongoing investigation. Several limitations—such as the geographic scope, reliance on sentiment as a proxy for conflict, and temporal constraints—highlight areas for further development. The next phases of this work will focus on extending the analysis to the full legislative archive spanning from 1981 to the present, incorporating manual annotation of conflict to replace or complement sentiment analysis, and experimenting with additional features and more advanced models, including large language models and neural networks. We also aim to validate the generalizability of our models using data from other regions and, most critically, to translate analytical results into concrete policy recommendations—ultimately the core objective of this research.

By addressing these next steps, we aim to deepen the theoretical understanding of political conflict and contribute practical tools for fostering transparency, fairness, and inclusivity in decision-making. This integrative and ethical approach to predictive analytics holds promise not only for political institutions but for any domain where conflict, representation, and bias intersect.

Acknowledgments

Special thanks to Consuelo R. Nava and Stefano Tedeschi for their kind support and valuable help throughout the development of this work.

Funded by the project “Gender Inclusion and Artificial Intelligence for Conflict Prediction” as part of the “2024 Ordinary Grants Call” – first session issued by Fondazione CRT – CUP B67G24000380009.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] G. Babaei, P. Giudici, E. Raffinetti, Safeaipackage: a python package for ai risk measurement, Available at SSRN (2024).

[2] R. Kreitner, A. Kinicki, Organizational Behavior, McGraw Hill Irwin., 2008.

[3] S. L. McShane, S. L. Steen, K. Tasa, Canadian organizational behaviour, McGraw-Hill Ryerson Toronto, ON, Canada, 2004.

[4] M. A. Rahim, Managing conflict in organizations, Routledge, 2023.

[5] K. W. Thomas, R. H. Kilmann, Comparison of four instruments measuring conflict behavior, Psychological reports 42 (1978) 1139–1145.

[6] M. Caprioli, Primed for violence: The role of gender inequality in predicting internal conflict, International studies quarterly 49 (2005) 161–178.
[7] E. Melander, Gender equality and intrastate armed conflict, International Studies Quarterly 49 (2005) 695–714.
[8] J. Gray, Men are from Mars, women are from Venus, HarperCollins, 1993.
[9] H. Hegre, J. Karlsen, H. M. Nygård, H. Strand, H. Urdal, Predicting armed conflict, 2010–2050, International Studies Quarterly 57 (2013) 250–270.
[10] H. Hegre, H. M. Nygård, P. Landsverk, Can we predict armed conflict? how the first 9 years of published forecasts stand up to reality, International Studies Quarterly 65 (2021) 660–668.
[11] C. Perry, Machine learning and conflict prediction: a use case, Stability: International Journal of Security and Development 2 (2013) 56.
[12] F. Ettensperger, Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field, Quality & Quantity 54 (2020) 567–601.
[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
[14] C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, ACL, 2019, pp. 380–385. URL: https://aclanthology.org/N19-1035/.