

PAN 2026

Lab on Digital Text Forensics and Stylometry



pan.webis.de



pan@webis.de

Henry Plutz¹

Andreas Jakoby¹

Janek Bevendorff¹

Martin Potthast^{2, 3, 4}

Benno Stein¹

¹ Bauhaus-Universität Weimar

² University of Kassel

³ hessian.AI

⁴ ScaDS.AI

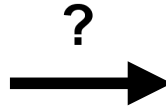
Overview: Text Watermarking

Text Watermarking

Motivation



Politician A

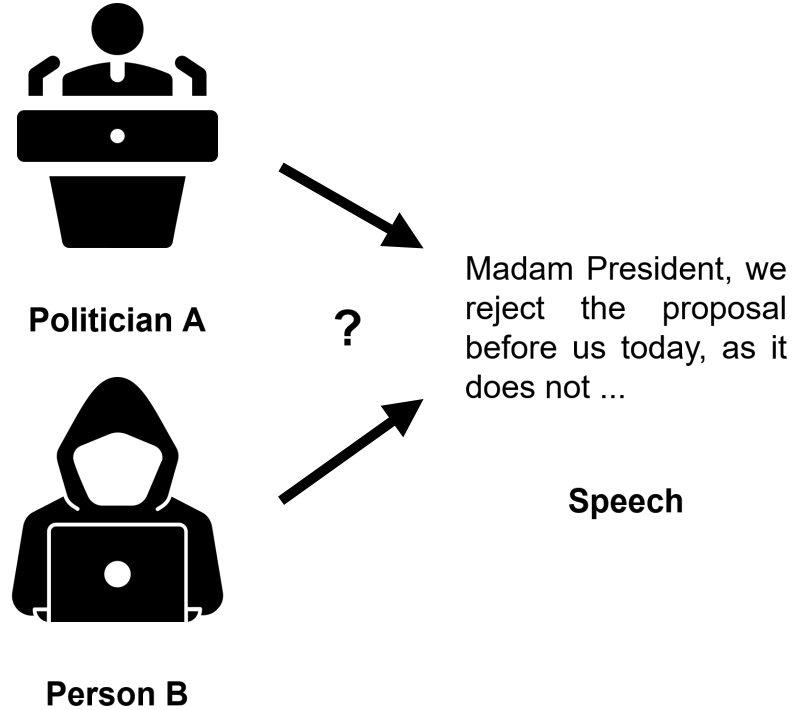


Madam President, we
reject the proposal
before us today, as it
does not ...

Speech

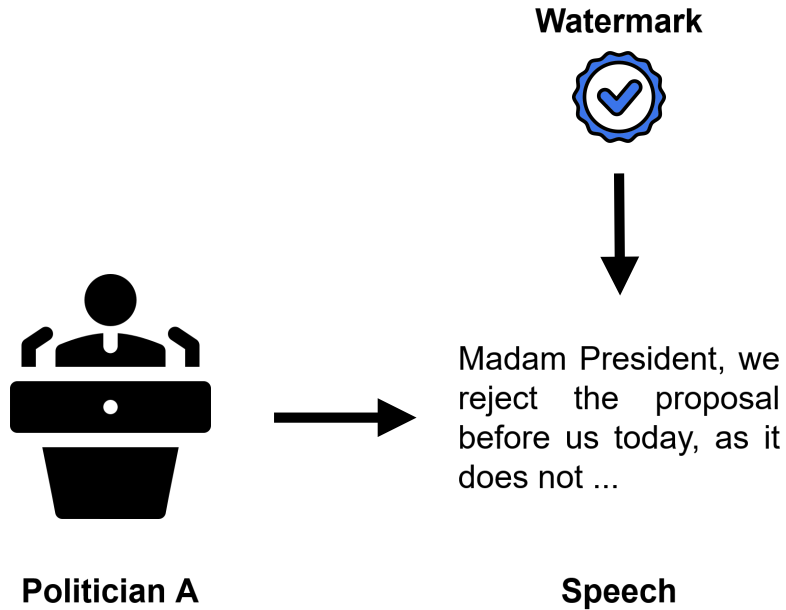
Text Watermarking

Motivation



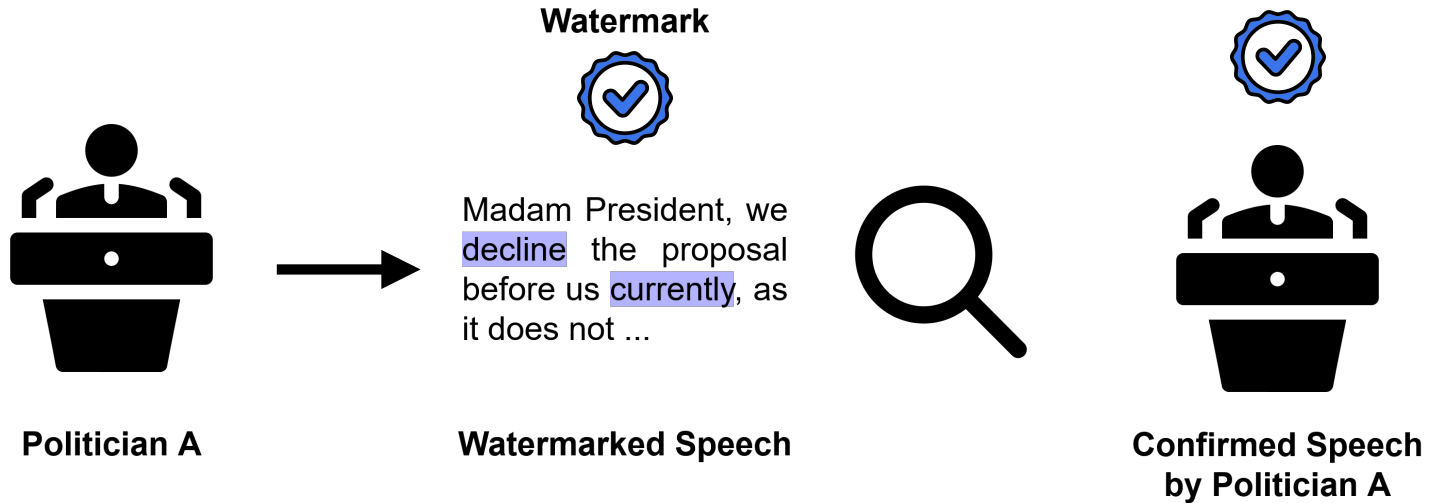
Text Watermarking

Motivation



Text Watermarking

Motivation



Text Watermarking @ PAN'26

The Text Watermarking Task @ PAN'26 aims to achieve the following:

- ❑ First ever shared task on Text Watermarking
- ❑ Benchmark the robustness of different watermarking systems against text obfuscations
- ❑ Advance evaluation of text watermarking methods

Text Watermarking

Post-Hoc vs. Generation-Inherent Watermarking

- Post-Hoc Watermarking:



Text

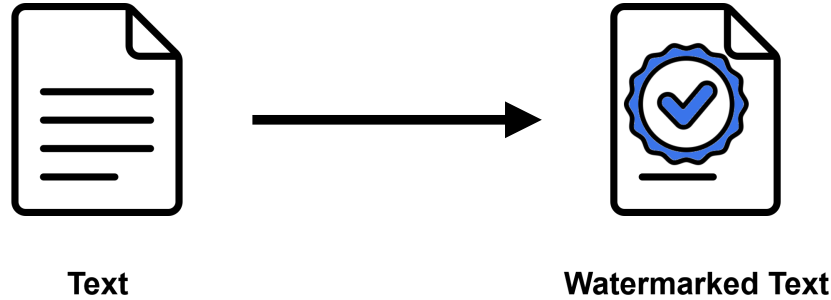


Watermarked Text

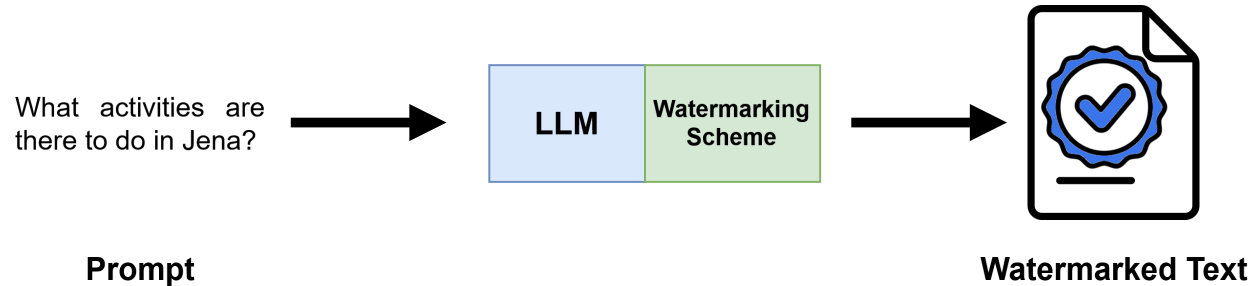
Text Watermarking

Post-Hoc vs. Generation-Inherent Watermarking

- Post-Hoc Watermarking:



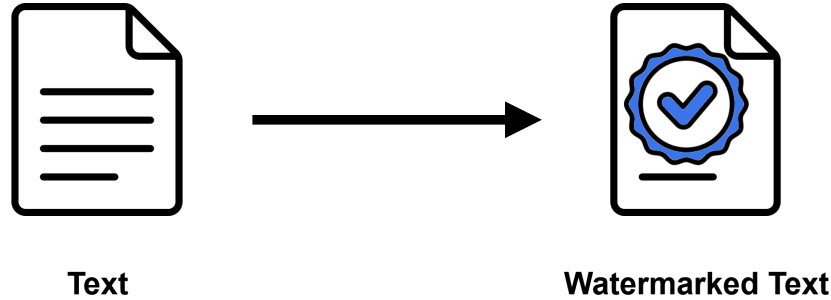
- Generation-Inherent Watermarking:



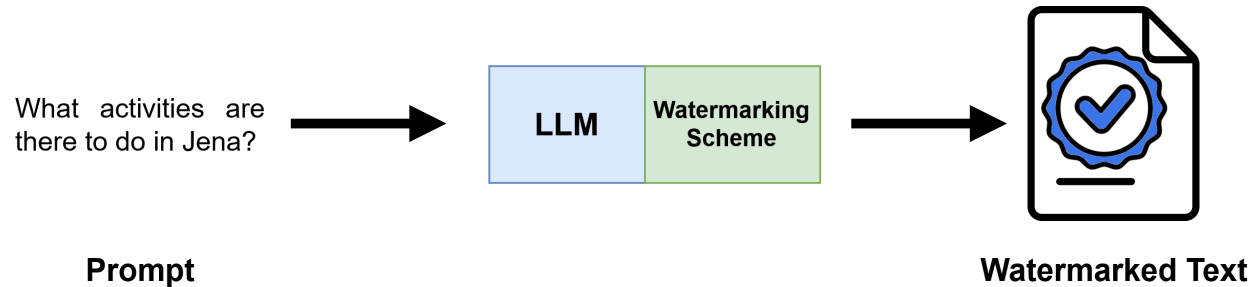
Text Watermarking

Post-Hoc vs. Generation-Inherent Watermarking

- ❑ **Post-Hoc Watermarking: Focus of this shared task!**



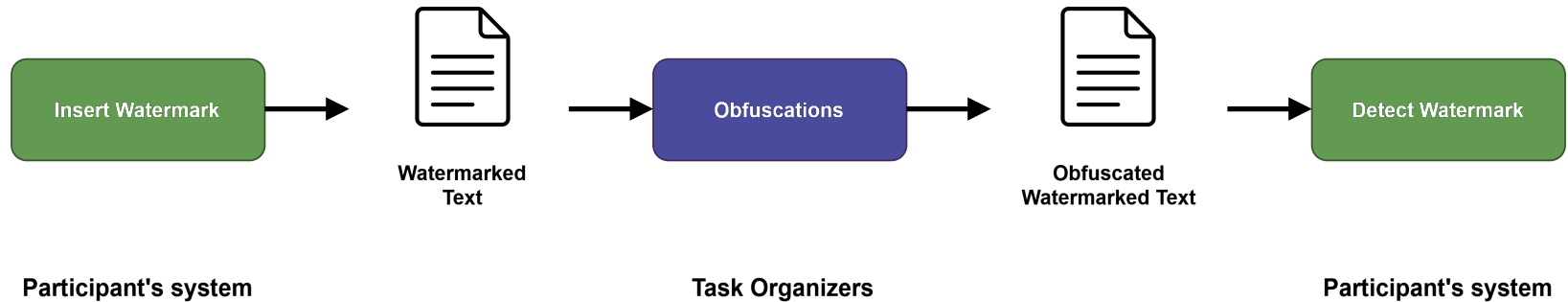
- ❑ **Generation-Inherent Watermarking:**



Text Watermarking

Task Formulation

1. Given a text, insert a watermark into that text.
2. We run the submitted watermarking system on the test dataset and obfuscate the resulting texts.
3. Detect the watermark in the obfuscated texts.



Text Watermarking

Dataset

- ❑ Political speeches from the European Parliament (ParlLawSpeech [Schwalbach et al., 2025])
- ❑ 300 speeches in the test dataset
- ❑ English language
- ❑ Dataset contains texts of different lengths (ranging from 50 – 1200 words)
- ❑ Dataset was cleaned to remove meta-comments and other textual artefacts

Text Watermarking

Obfuscations

- ❑ **Typo Attack:** Randomly replace one letter in words with adjacent letters on a QWERTY-keyboard.
- ❑ **Paraphrasing Attack:** Paraphrase sentences using another LLM.
- ❑ **Copy-Paste Attack:** Replace watermarked sentences with unwatermarked sentences.

Text Watermarking

Obfuscations

Text

--

Sentence Split

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Text Watermarking

Obfuscations

Text



Sentence Split



Level 1



Level 2



Level 3



Level 4

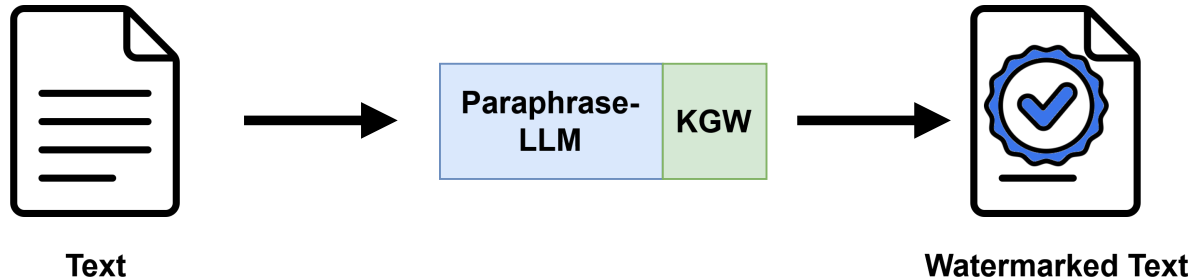


Four levels of obfuscation severities

Text Watermarking

Baseline

- ❑ KGW [Kirchenbauer et al., 2023] watermark implemented for post-hoc watermarking
- ❑ In general a generation-inherent watermarking method
- ❑ Here: Paraphrases and inserts the watermark during generation



Text Watermarking

Submission Evaluation

- ❑ Score each text individually by TWF and average the score on all texts.
- ❑ Combines watermark detection performance and similarity of the watermarked text to the original text.

$$\text{TWF} = \max(\text{BLEU}, \text{BERTScore}) \cdot \text{BalancedAccuracy}$$

Text Watermarking

Submission Evaluation

- ❑ Score each text individually by TWF and average the score on all texts.
- ❑ Combines watermark detection performance and similarity of the watermarked text to the original text.

$$\text{TWF} = \max(\text{BLEU}, \text{BERTScore}) \cdot \text{BalancedAccuracy}$$

- ❑ Syntactic Similarity

Text Watermarking

Submission Evaluation

- ❑ Score each text individually by TWF and average the score on all texts.
- ❑ Combines watermark detection performance and similarity of the watermarked text to the original text.

$$\text{TWF} = \max(\text{BLEU}, \text{BERTScore}) \cdot \text{BalancedAccuracy}$$

- ❑ Syntactic Similarity
- ❑ Semantic Similarity

Text Watermarking

Submission Evaluation

- ❑ Score each text individually by TWF and average the score on all texts.
- ❑ Combines watermark detection performance and similarity of the watermarked text to the original text.

$$\text{TWF} = \max(\text{BLEU}, \text{BERTScore}) \cdot \text{BalancedAccuracy}$$

- ❑ Syntactic Similarity
- ❑ Semantic Similarity
- ❑ Detection Performance

Text Watermarking

Submissions

- ❑ Di Nuovo et al.: Multi-channel system, lexical substitutions, syntactic transformations
- ❑ Ibrahim: Combination and Extension of KGW and PostMark
- ❑ Elshamy et al. (1): Improved Baseline parameters
- ❑ Elshamy et al. (2): KGW for individual sentences
- ❑ Rudolph and Stadelmann: Two-channel system, SimMark, KGW

Text Watermarking

Submissions Ranking

* No notebook submitted

Team	Authors	TWF↓	BA	BLEU	BERTScore
<i>yeyang</i> *	-	0.818	0.824	0.948	0.992
aimoment (Single-key)	Ibrahim	0.775	0.785	0.904	0.988
team	Rudolph and Stadelmann	0.762	0.773	0.857	0.987
JRC-GENESIS (v2)	Di Nuovo et al.	0.761	0.785	0.642	0.969
aimoment (Dual-key)	Ibrahim	0.734	0.753	0.789	0.975
<i>alex-csed-3n</i> *	-	0.727	0.819	0.071	0.887
JRC-GENESIS (v1)	Di Nuovo et al.	0.686	0.720	0.572	0.952
KGW Baseline		0.633	0.720	0.014	0.880

Text Watermarking

Conclusion and Outlook

❑ **Conclusion:**

- All teams beat the baseline
- Multi-channel watermarking systems worked well on the test set

❑ **Outlook:**

- Text Watermarking Task will be present @ PAN'27
- Bigger dataset
- Wider range of attacks

PAN 2026

Lab on Digital Text Forensics and Stylometry



pan.webis.de



pan@webis.de

Thanks!



[Task Website](#)



[GitHub Repository](#)