

## A. Features used in our implementation

**Table 1**

The 25 features used in our implementation of the superiority theory classifier using the THInC framework.

Computational tool	Time series	Manifestation	Derived proxy features
TweetNLP offensive language identification [1]	Offense, subsequence-based	Increasing offensive language towards others.	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Linear fit standard error</li> <li>- Skewness</li> <li>- Symmetry looking</li> <li>- Aggregated linear trend</li> <li>- Ratio of crossings at 0.9</li> <li>- Ratio of crossings at 0.5</li> </ul>
Author stance detection model [2]	Attack, subsequence-based	Increasing aggression or confrontation, enhancing the group cohesion of those who align against the target	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Linear fit standard error</li> <li>- Skewness</li> <li>- Symmetry looking</li> <li>- Aggregated linear trend</li> <li>- Ratio of crossings at 0.9</li> <li>- Ratio of crossings at 0.5</li> </ul>
Hate detection model [3]	Hate, subsequence-based	Increasing hate speech towards perceived inferiors	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Linear fit standard error</li> <li>- Skewness</li> <li>- Symmetry looking</li> <li>- Aggregated linear trend</li> <li>- Ratio of crossings at 0.9</li> <li>- Ratio of crossings at 0.5</li> </ul>
TweetNLP sentiment analysis [1]	Neutrality, subsequence-based	Absence of neutrality	<ul style="list-style-type: none"> <li>- Absolute energy</li> <li>- Mean absolute change</li> </ul>
TweetNLP sentiment analysis [1]	Positivity, subsequence-based	No consistent positivity (which would not align with superiority)	<ul style="list-style-type: none"> <li>- Large standard deviation</li> </ul>
TweetNLP sentiment analysis [1]	Negativity, subsequence-based	No consistent negativity (this would be no humor, just negativity)	<ul style="list-style-type: none"> <li>- Large standard deviation</li> </ul>

**Table 2**

The 48 features used in our implementation of the incongruity theory classifier using the THInC framework.

Computational tool	Time series	Manifestation	Derived proxy features
Probabilities outputted by Llama 2 [4]	Llama probabilities, token-based	Disruption in normal expectancy shown by sudden changes in probabilities	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE (complexity estimate)</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP sentiment analysis [1]	Positivity, subsequence-based	Sudden spikes in positivity in contexts where it is not anticipated	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP sentiment analysis [1]	Negativity, subsequence-based	Unexpected drops or increases in negativity that counteract the expected emotional tone	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP emotion recognition [1]	Joy, subsequence-based	Bursts of joy that break from the narrative or emotional flow	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP emotion recognition [1]	Optimism, subsequence-based	Abrupt transitions to optimism in seemingly unsuitable or unexpected scenarios	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP emotion recognition [1]	Sadness, subsequence-based	Placement of sadness in a context that typically calls for happiness	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
TweetNLP emotion recognition [1]	Anger, subsequence-based	Unexpected surges of anger that contradict the prevailing mood or setting	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>
Subjective bias detection model [5]	Subjectivity, subsequence-based	Shifts in subjectivity that unexpectedly skew perception in otherwise neutral or objective narratives	<ul style="list-style-type: none"> <li>- Max change</li> <li>- CID CE</li> <li>- Ratio of crossings at 0.5</li> <li>- Ratio of wavelet peaks</li> <li>- Ratio of peaks</li> <li>- Ratio beyond 2 sigma</li> </ul>

**Table 3**

The 46 features used in our implementation of the relief theory classifier using the THInC framework.

Computational tool	Time series	Manifestation	Derived proxy features
TweetNLP emotion recognition [1]	Optimism, subsequence-based	Gradual increase in optimism as tension is released	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
TweetNLP emotion recognition [1]	Joy, subsequence-based	Gradual increase in joy as tension is released	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
TweetNLP emotion recognition [1]	Anger, subsequence-based	Decrease in anger as tension is released	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
TweetNLP emotion recognition [1]	Sadness, subsequence-based	Reduction in sadness as tension is released	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
Hate detection model [3]	Hate, subsequence-based	Decrease in hate speech as tension is released	<ul style="list-style-type: none"> <li>- Linear fit standard error</li> <li>- Aggregate linear trend</li> <li>- Ratio of crossings at 0.5</li> <li>- Skewness</li> <li>- Symmetry looking</li> <li>- First location of maximum/minimum</li> <li>- Energy ratio by chunks (segment 3 of 3)</li> <li>- Mass center</li> </ul>
Adult language detection model [6]	Adult language, subsequence-based	Change in the use of adult language as tension is released	<ul style="list-style-type: none"> <li>- Linear fit slope/standard error</li> <li>- Skewness</li> <li>- Symmetry looking</li> <li>- First location of maximum/minimum</li> <li>- Energy ratio by chunks (segment 0 of 3)</li> <li>- Mass center</li> <li>- Mean second derivative</li> </ul>

**Table 4**

The 36 features used in our implementation of the surprise disambiguation model classifier using the THInC framework.

Computational tool	Time series	Manifestation	Derived proxy features
Probabilities outputted by Llama 2 [4]	Llama probabilities, token-based	Unexpected twist during the resolution of a joke	<ul style="list-style-type: none"> <li>- Energy ratio by chunks (segment 2 of 2)</li> <li>- Mass center</li> <li>- Linear fit standard error</li> <li>- Ratio beyond 1 sigma</li> <li>- Maximum change</li> <li>- Maximum change timing</li> </ul>
TweetNLP offensive language identification [1]	Offense, subsequence-based	Resolution of a joke gives a change in offense	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative central</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
Author stance detection model [2]	Attack, subsequence-based	Resolution of a joke gives a change in attack	<ul style="list-style-type: none"> <li>- Linear fit slope</li> <li>- Mean second derivative central</li> <li>- Energy ratio by chunks (segments 0 and 3 of 3)</li> <li>- Mass center</li> <li>- Skewness</li> <li>- Symmetry looking</li> </ul>
ConceptNet to retrieve related words + GloVe word embeddings to compute the pairwise distances between these words [7, 8]	Ambiguity, token-based	Ambiguity in jokes is built up and then clarified	<ul style="list-style-type: none"> <li>- Mass center</li> <li>- Mass 25th percentile</li> <li>- Skewness</li> <li>- Linear fit slope</li> <li>- Linear fit standard error</li> <li>- Aggregated linear trend</li> </ul>
Confidence in the top part-of-speech by the spaCy tagger [9]	Morphosyntactic ambiguity, token-based	Morphosyntactic ambiguity plays with language rules to create and resolve incongruities.	<ul style="list-style-type: none"> <li>- Mass center</li> <li>- Mass at 25th percentile</li> <li>- Skewness</li> <li>- Linear fit slope</li> <li>- Linear fit standard error</li> <li>- Aggregated linear trend</li> </ul>
Adult language detection model [6]	Adult language, subsequence-based	An introduction of adult language can be the resolution of an incongruity.	<ul style="list-style-type: none"> <li>- First location of maximum</li> <li>- First location of minimum</li> <li>- Mass center</li> <li>- Energy ratio by chunks (segment 2 of 2)</li> </ul>

## References

- [1] J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa-Anke, F. Liu, E. Martínez-Cámara, et al., TweetNLP: Cutting-Edge Natural Language Processing for Social Media, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Abu Dhabi, U.A.E., 2022.
- [2] Gajewska E., Stance-Tw (revision cc81e6d), 2023. <https://huggingface.co/eevvgg/Stance-Tw>.
- [3] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, Learning from the worst: Dynamically generated datasets to improve online hate detection, 2021. [arXiv:2012.15761](https://arxiv.org/abs/2012.15761).
- [4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, [arXiv preprint arXiv:2307.09288](https://arxiv.org/abs/2307.09288) (2023).
- [5] C. F. F. Labs, Text style transfer - neutralizing subjectivity bias with huggingface transformers, <https://github.com/fastforwardlabs/text-style-transfer>, 2022.
- [6] Valurank, finetuned-distilbert-adult-content-detection (revision 5383ff5), 2022. URL: <https://huggingface.co/valurank/finetuned-distilbert-adult-content-detection>.
- [7] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, 2017. URL: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972>.
- [8] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>. doi:10.3115/v1/D14-1162.
- [9] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). doi:10.5281/zenodo.1212303.