



Universiteit  
Leiden  
The Netherlands

## Opinion diversity through hybrid intelligence

Meer, M.T. van der

### Citation

Meer, M. T. van der. (2025, March 26). *Opinion diversity through hybrid intelligence*. *SIKS Dissertation Series*. Retrieved from <https://hdl.handle.net/1887/4209024>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4209024>

**Note:** To cite this publication please use the final published version (if applicable).

# Contents

<b>Summary</b>	<b>vii</b>
<b>Samenvatting</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Questions . . . . .	4
1.2 Research Methodology . . . . .	5
1.2.1 Fundamental Issues (Q1) . . . . .	5
1.2.2 Hybrid Intelligence (Q2) . . . . .	6
1.2.3 Perspective Hierarchy (Q3) . . . . .	9
1.3 Dissertation Scope . . . . .	11
1.4 Outlook . . . . .	12
<b>I NLP for Online Discussion Analysis</b>	<b>15</b>
<b>2 An Empirical Analysis of Diversity in Argument Summarization</b>	<b>19</b>
2.1 Introduction . . . . .	20
2.2 Related Work . . . . .	21
2.2.1 Key Point Analysis . . . . .	21
2.2.2 Opinion Summarization . . . . .	21
2.2.3 Diversity in Societal Decision Making . . . . .	21
2.3 Method . . . . .	21
2.3.1 Task setup . . . . .	22
2.3.2 Modeling Diversity in Key Point Analysis . . . . .	22
2.4 Experimental Setup . . . . .	23
2.4.1 Data . . . . .	23
2.4.2 Approaches . . . . .	24
2.4.3 Evaluation Metrics . . . . .	25
2.5 Results and Discussion . . . . .	26
2.5.1 KPG Performance . . . . .	26
2.5.2 KPM Performance . . . . .	26
2.5.3 Analysis . . . . .	27
2.6 Conclusion . . . . .	30
<b>3 Will It Blend? Mixing Training Paradigms &amp; Prompting for Argument Quality Prediction</b>	<b>33</b>
3.1 Introduction . . . . .	34
3.2 Related Work . . . . .	34
3.3 Data and Training Paradigms . . . . .	35
3.3.1 Data . . . . .	35

3.3.2	Training Paradigms . . . . .	35
3.4	Approach. . . . .	36
3.4.1	Implementation details . . . . .	37
3.5	Experiments and Results . . . . .	37
3.5.1	Error Analysis . . . . .	38
3.6	Conclusion . . . . .	39
3.7	Access and Responsible Research . . . . .	39
<b>II</b>	<b>Hybrid Intelligence for NLP</b>	<b>41</b>
<b>4</b>	<b>A Hybrid Intelligence Method for Argument Mining</b>	<b>45</b>
4.1	Introduction . . . . .	46
4.2	Related work . . . . .	48
4.2.1	Computational Argument Analysis . . . . .	48
4.2.2	Summarization of Arguments. . . . .	49
4.3	Method . . . . .	49
4.3.1	Opinion Corpora . . . . .	50
4.3.2	Key Argument Annotation . . . . .	51
4.3.3	Key Argument Consolidation. . . . .	52
4.3.4	Key Argument Selection . . . . .	54
4.4	Experimental Setup. . . . .	54
4.4.1	Phase 1: Key Argument Annotation. . . . .	55
4.4.2	Phase 2: Key Argument Consolidation . . . . .	56
4.4.3	Phase 3: Key Argument Selection . . . . .	56
4.4.4	Baselines . . . . .	58
4.5	Results . . . . .	60
4.5.1	Annotator Agreement . . . . .	60
4.5.2	Phase 1: Key Argument Annotation. . . . .	61
4.5.3	Phase 2: Key Argument Consolidation . . . . .	63
4.5.4	Phase 3: Key Argument Selection . . . . .	64
4.5.5	Comparison with Automated Baseline . . . . .	67
4.5.6	Comparison with Manual Baseline . . . . .	68
4.6	Discussion . . . . .	69
4.7	Conclusion and Future Directions . . . . .	71
<b>5</b>	<b>Annotator-Centric Active Learning for Subjective NLP Tasks</b>	<b>73</b>
5.1	Introduction . . . . .	74
5.2	Related work . . . . .	75
5.2.1	Learning with annotator disagreement . . . . .	75
5.2.2	Active Learning . . . . .	75
5.3	Method . . . . .	76
5.3.1	Soft-label prediction . . . . .	76
5.3.2	Active Learning . . . . .	76
5.3.3	Annotator-Centric Active Learning. . . . .	77

5.4	Experimental Setup . . . . .	78
5.4.1	Datasets . . . . .	78
5.4.2	Evaluation metrics . . . . .	78
5.4.3	Training procedure . . . . .	79
5.5	Results . . . . .	79
5.5.1	Highlights . . . . .	79
5.5.2	Efficiency and Fairness . . . . .	80
5.5.3	Convergence . . . . .	83
5.5.4	Impact of subjectivity. . . . .	83
5.6	Conclusion . . . . .	85
<b>III Social Science with Hybrid Intelligence</b>		<b>87</b>
<b>6</b>	<b>Do Differences in Values Influence Disagreements in Online Discussions?</b>	<b>91</b>
6.1	Introduction . . . . .	92
6.2	Related Work. . . . .	93
6.2.1	(Dis)-agreement and discussion analysis . . . . .	93
6.2.2	Value models . . . . .	94
6.2.3	Value estimation . . . . .	94
6.3	Method . . . . .	94
6.3.1	Data . . . . .	95
6.3.2	Value Extraction . . . . .	95
6.3.3	Value Profile Estimation . . . . .	96
6.4	Experiments and Results . . . . .	96
6.4.1	Training Models for Value Estimation. . . . .	96
6.4.2	Value Profile Estimation . . . . .	97
6.4.3	Value Conflicts and Disagreement . . . . .	98
6.4.4	Use Case: Predicting (Dis-)agreement. . . . .	102
6.5	Conclusion . . . . .	103
<b>IV Conclusions</b>		<b>105</b>
<b>7</b>	<b>Contributions and Future Work</b>	<b>107</b>
7.1	Research Findings . . . . .	108
7.1.1	NLP for Perspective Analysis . . . . .	109
7.1.2	Hybrid Intelligence for NLP . . . . .	110
7.1.3	Perspective Hierarchy. . . . .	112
7.2	Contributions . . . . .	112
7.2.1	Scientific Relevance. . . . .	112
7.2.2	Societal Relevance . . . . .	114
7.3	Limitations. . . . .	115
7.4	Future Work . . . . .	116

<b>V</b>	<b>Appendices</b>	<b>119</b>
<b>A</b>	<b>An Empirical Analysis of Diversity in Argument Summarization</b>	<b>121</b>
A.1	Detailed Experimental Setup . . . . .	121
A.1.1	Data . . . . .	121
A.1.2	Per-approach Specifics . . . . .	122
A.1.3	Evaluation metrics . . . . .	124
A.2	Additional results. . . . .	125
A.2.1	Detailed ROUGE scores for Key Point Generation. . . . .	125
A.2.2	Additional BERTScores for Key Point Generation . . . . .	125
A.2.3	Long-tail experiment for KPG . . . . .	125
A.2.4	ChatGPT generated key points for PVE . . . . .	126
<b>B</b>	<b>Will It Blend? Mixing Training Paradigms &amp; Prompting for Argument Quality Prediction</b>	<b>129</b>
B.1	Hyperparameters. . . . .	129
B.2	Additional results. . . . .	130
B.2.1	Per-label Performance . . . . .	130
B.2.2	Label confusion . . . . .	130
B.2.3	Seed Variance . . . . .	130
B.2.4	Topics . . . . .	131
<b>C</b>	<b>A Hybrid Intelligence Method for Argument Mining</b>	<b>133</b>
C.1	Experiment Protocol & Description . . . . .	133
C.1.1	Preliminaries . . . . .	133
C.1.2	Phase 1: Argument Annotation . . . . .	133
C.1.3	Phase 2: Argument Consolidation . . . . .	134
C.1.4	Comparison to Automated Baseline. . . . .	134
C.1.5	Annotation platform . . . . .	134
C.2	Method Details . . . . .	135
C.2.1	Parallel Pairwise Annotation Algorithm. . . . .	135
C.2.2	Hyperparameters. . . . .	135
C.3	Detailed Results . . . . .	138
C.3.1	Unclear Translation Actions . . . . .	138
C.3.2	Clustering Arguments . . . . .	138
C.3.3	Key Arguments. . . . .	138
<b>D</b>	<b>Annotator-Centric Active Learning for Subjective NLP Tasks</b>	<b>149</b>
D.1	Detailed Experimental Setup . . . . .	149
D.1.1	Dataset details . . . . .	149
D.1.2	Hyperparameters. . . . .	149
D.1.3	Training details. . . . .	150
D.1.4	ACAL annotator strategy details . . . . .	150
D.1.5	Disagreement rates . . . . .	151
D.2	Detailed results overview . . . . .	152
D.2.1	Annotator-Centric evaluation for other MFTC and MHS tasks . . . . .	152
D.2.2	Training process . . . . .	152

<b>E</b>	<b>Do Differences in Values Influence Disagreements in Online Discussions?</b>	<b>159</b>
E.1	Methodological details . . . . .	159
E.1.1	Training Value extraction methods . . . . .	159
E.1.2	Annotator experiment . . . . .	160
E.1.3	Training agreement analysis models. . . . .	162
E.2	Additional Results . . . . .	165
E.2.1	Value Extraction . . . . .	165
E.2.2	Value Survey . . . . .	165
E.2.3	Qualitative Examples of Value Conflicts and (Dis-)agreement . . . . .	165
E.2.4	Decomposition of $BF_{10}$ results . . . . .	165
E.2.5	Kendall $\tau$ vs. Spearman $\rho$ . . . . .	166
E.2.6	Agreement Analysis . . . . .	166
	<b>Bibliography</b>	<b>171</b>
	<b>Acknowledgments</b>	<b>217</b>
	<b>Curriculum Vitæ</b>	<b>219</b>
	<b>List of Publications</b>	<b>221</b>
	<b>SIKS Dissertations</b>	<b>223</b>

