COMP0173: Coursework 2

Arina Bekenova

December 11, 2025

1 Baseline Results Replication: Comparison With HEARTS Paper

Table 1: Comparison of ALBERT-V2 model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	**Original**	**Replicated**	**Diff. (pp)**	**Within 5%**
MGSD → MGSD	79.7%	78.1%	1.6	Yes
MGSD → AWinoQueer	74.7%	60.3%	14.3	No
MGSD → ASeeGULL	75.9%	73.6%	2.3	Yes
MGSD → EMGSD	79.3%	77.2%	2.1	Yes
AWinoQueer → MGSD	60.0%	61.4%	1.4	Yes
AWinoQueer → AWinoQueer	97.3%	97.9%	0.6	Yes
AWinoQueer → ASeeGULL	70.7%	74.2%	3.5	Yes
AWinoQueer → EMGSD	62.8%	64.4%	1.6	Yes
ASeeGULL → MGSD	63.1%	64.1%	1	Yes
ASeeGULL → AWinoQueer	66.8%	75.8%	9	No
ASeeGULL → ASeeGULL	88.4%	87.8%	0.6	Yes
ASeeGULL → EMGSD	64.5%	65.7%	1.2	Yes
EMGSD → MGSD	80.2%	79.4%	0.8	Yes
EMGSD → AWinoQueer	97.4%	97.4%	0	Yes
EMGSD → ASeeGULL	87.3%	88.5%	1.2	Yes
EMGSD → EMGSD	81.5%	80.8%	0.7	Yes

2 Adapted Models Performance

=

Table 2: Comparison of BERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	**Original**	**Replicated**	**Diff. (pp)**	**Within 5%**
MGSD → MGSD	81.2%	80.9%	0.3	Yes
MGSD → AWinoQueer	77.9%	78.3%	0.4	Yes
MGSD → ASeeGULL	69.9%	71.1%	1.2	Yes
MGSD → EMGSD	80.6%	80.4%	0.2	Yes
AWinoQueer → MGSD	59.1%	62.7%	3.6	Yes
AWinoQueer → AWinoQueer	97.9%	99.1%	1.2	Yes
AWinoQueer → ASeeGULL	72.5%	72.7%	0.2	Yes
AWinoQueer → EMGSD	62.3%	65.1%	2.8	Yes
ASeeGULL → MGSD	61.0%	63.1%	2.1	Yes
ASeeGULL → AWinoQueer	78.6%	82.3%	3.7	Yes
ASeeGULL → ASeeGULL	89.6%	88.9%	0.7	Yes
ASeeGULL → EMGSD	63.3%	65.3%	2	Yes
EMGSD → MGSD	81.7%	81.3%	0.4	Yes
EMGSD → AWinoQueer	97.6%	97.6%	0	Yes
EMGSD → ASeeGULL	88.9%	89.1%	1.2	Yes
EMGSD → EMGSD	82.8%	82.5%	0.3	Yes

Table 3: Comparison of DistilBERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	**Original**	**Replicated**	**Diff. (pp)**	**Within 5%**
MGSD → MGSD	78.3%	78.3 %	0	Yes
MGSD → AWinoQueer	75.6%	75.3%	0.3	Yes
MGSD → ASeeGULL	73.0%	75.3%	2.3	Yes
MGSD → EMGSD	78.0%	78.1%	0.1	Yes
AWinoQueer → MGSD	61.1%	62.2%	1.1	Yes
AWinoQueer → AWinoQueer	98.1%	97.4%	0.7	Yes
AWinoQueer → ASeeGULL	72.1%	70.1%	2	Yes
AWinoQueer → EMGSD	64%	64.8%	0.8	Yes
ASeeGULL → MGSD	62.7%	62.1%	0.6	Yes
ASeeGULL → AWinoQueer	82.1%	86.8%	4.7	Yes
ASeeGULL → ASeeGULL	89.8%	88.3%	1.5	Yes
ASeeGULL → EMGSD	65.1%	64.8%	0.3	Yes
EMGSD → MGSD	79.0%	79.0%	0	Yes
EMGSD → AWinoQueer	98.8%	98.4%	0.4	Yes
EMGSD → ASeeGULL	91.9%	90.8%	1.1	Yes
EMGSD → EMGSD	80.6%	80.5%	0.1	Yes

Table 4: Comparison of Adapted Model Macro F1 scores (%)

Model Type	**Emissions**	**Training Data**	**Test Set Macro F1 Score%**
LR-TFIDF	0.01g	RBS	93.1%
LR-TFIDF	0.01g	RBSA	51.8%
LR-Embeddings	0.21g	RBS	92.4%
LR-Embeddings	0.15g	RBSA	65.5%
DeepPavlov-RuBert	1.28g	RBS	97.1%
DeepPavlov-RuBert	1.09g	RBSA	72.4%
AI-Forever-RuBert	1.36g	RBS	97.2%
AI-Forever-RuBert	1.08g	RBSA	**74.3%**
XLM-RoBERTa	Unknown	RBS	**97.9%**
XLM-RoBERTa	Unknown	RBSA	73.8%