

20 Newsgroups 文本分类评估报告

本报告由 `evaluation/evaluate_models.py` 与 `evaluation/generate_report.py` 自动生成。

1. 数据集说明

数据集为 **20 Newsgroups**，共 20 个主题类别，测试集样本数为 **7532**。

本项目数据加载函数为 `data/data_loader.py`，使用 `fetch_20newsgroups` 并移除 `headers/footers/quotes` 噪声字段。

2. Baseline 与模型说明

- tfidf_logreg**: `TfidfVectorizer + LogisticRegression`，模型工件路径见 `modeling/configs/run_tfidf_logreg_metadata.json`。
- llm_classifier**: 基于 Hugging Face 因果语言模型的 zero-shot 分类（当前记录模型：`Qwen/Qwen2.5-1.5B-Instruct`）。

3. 配置说明

- 训练/推理入口: `main.py`
- 传统基线配置: `modeling/configs/run_tfidf_logreg_metadata.json`
- LLM 运行配置: `modeling/configs/llm_run_metadata.json`
- 评估输出目录: `evaluation/outputs`

4. 全流程结构 (Pipeline)

- 加载 20 Newsgroups 数据（训练/测试）
- 训练 TF-IDF + Logistic Regression 基线并保存模型
- 运行 LLM zero-shot 预测并保存 JSONL 结果
- 统一评估脚本读取两种模型输出，计算宏平均与按类指标
- 生成评估工件（JSON/CSV）并产出中文 PDF 报告

5. 输出说明

- `evaluation/outputs/metrics_summary.json`：核心宏平均指标
- `evaluation/outputs/detailed_metrics.json`：按类指标与混淆矩阵
- `evaluation/outputs/per_class_metrics.csv`：20 类别逐类 Precision/Recall/F1
- `evaluation/outputs/confusion_matrix_*.csv`：混淆矩阵数据

6. 评估指标选择说明

本项目使用 **Macro-Precision / Macro-Recall / Macro-F1** 作为核心指标。原因：20 类别任务中，不同类别难度差异明显，宏平均能够让每个类别等权重参与评估，避免被高频类别主导。

模型	Macro-Precision	Macro-Recall	Macro-F1	未知预测占比
tfidf_logreg	0.6653	0.6469	0.6450	0.0000
llm_classifier	0.3580	0.1865	0.2001	0.2775

7. 混淆矩阵与误差分析

TF-IDF + LogReg 混淆矩阵 (20x20)

真实 \预测	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motocycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
alt.atheism	30	3	1	0	4	3	0	3	13	16	1	2	5	10	16	64	14	10	6	9
comp.graphics	51	20	10	12	21	8	4	6	9	0	3	14	3	13	1	2	0	0	0	0
comp.os.ms-windows.misc	28	236	33	15	14	0	2	5	23	0	4	5	4	12	1	3	2	2	1	1
comp.sys.ibm.pc.hardware	16	239	26	5	14	6	1	8	1	3	35	0	1	0	0	0	0	0	0	0
comp.sys.mac.hardware	5	32	246	2	8	13	4	18	3	1	26	1	7	1	0	0	0	0	0	0
comp.windows.x	58	27	7	4	262	3	1	4	10	0	3	3	3	6	1	0	1	1	0	0
misc.forsale	3	3	15	12	0	311	11	5	10	1	1	8	1	4	1	2	1	1	0	0
rec.autos	3	2	0	2	3	12	265	18	34	2	2	23	6	6	1	8	1	4	1	1
rec.motocycles	5	0	0	2	0	7	28	298	23	0	2	11	3	7	3	7	0	1	0	0
rec.sport.baseball	1	0	1	3	5	1	8	326	22	0	4	2	1	5	0	3	4	0	0	0
rec.sport.hockey	0	0	0	1	0	1	5	31	341	1	1	1	3	3	2	0	1	1	1	1
sci.crypt	1	14	2	2	2	5	1	4	7	23	1	259	20	5	6	4	23	7	9	1
sci.electronics	22	12	24	16	7	12	15	9	14	2	13	222	10	12	0	2	0	0	1	1
sci.med	7	15	1	1	1	2	9	8	13	14	3	0	13	279	7	4	5	3	9	2
sci.space	6	9	3	0	2	0	3	6	6	23	1	3	17	12	289	1	3	3	8	0
soc.religion.christian	2	0	1	1	1	3	1	18	2	0	4	9	4	320	1	0	2	6	6	6
talk.politics.guns	2	1	2	0	2	6	9	19	3	9	4	6	8	8	251	4	16	6	6	6
talk.politics.mideast	2	0	1	0	1	3	9	20	1	2	3	3	3	11	7	273	12	1	1	1
talk.politics.misc	0	0	1	1	1	3	7	14	1	5	3	8	10	4	94	10	123	2	2	2
talk.religion.misc	1	1	0	2	2	5	8	11	2	1	0	10	5	71	22	8	9	39	39	39

LLM 混淆矩阵 (20x20, 仅统计可映射标签)

真实 \预测	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motocycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
alt.atheism	30	0	0	0	0	0	0	10	0	6	3	5	33	26	6	106	4	0	0	0
comp.graphics	50	3	8	2	1	19	53	0	2	0	0	89	0	1	19	4	0	1	0	0
comp.os.ms-windows.misc	1	5	12	62	3	8	11	51	0	9	1	89	1	1	35	1	0	4	0	0
comp.sys.ibm.pc.hardware	2	82	0	2	13	103	6	20	0	0	79	0	0	20	2	0	1	1	1	1
comp.sys.mac.hardware	5	48	16	1	13	98	4	18	0	1	55	0	0	17	1	0	1	2	2	2
comp.windows.x	2	18	18	2	9	16	36	0	7	0	2	95	3	0	20	3	0	2	1	1
misc.forsale	3	0	2	0	0	77	70	2	7	2	0	2	1	0	43	5	0	0	6	6
rec.autos	5	5	0	16	0	0	5	176	3	10	0	0	64	7	3	18	2	0	0	2
rec.motocycles	5	2	5	0	0	6	115	29	18	6	3	60	7	0	31	18	0	2	1	1
rec.sport.baseball	0	1	0	0	1	18	0	174	1	0	41	15	0	25	2	0	1	1	1	1
rec.sport.hockey	0	3	0	0	0	20	0	63	144	1	14	12	1	32	6	0	0	0	0	0
sci.crypt	2	2	1	1	0	0	2	5	0	4	0	114	86	6	2	56	9	0	1	0
sci.electronics	6	0	10	2	0	8	67	1	20	1	6	148	4	0	15	2	0	2	1	1
sci.med	2	0	0	1	0	0	1	25	3	24	3	0	49	158	0	25	2	0	0	1
sci.space	6	0	0	1	0	0	0	7	0	9	2	3	86	5	143	16	2	0	1	1
soc.religion.christian	0	0	0	0	0	12	4	7	0	26	53	58	1	117	2	0	0	3	3	3
talk.politics.guns	0	0	0	0	0	11	0	20	0	11	32	15	3	130	20	0	4	0	0	0
talk.politics.mideast	0	0	0	0	0	19	0	3	2	12	12	28	5	167	7	0	2	0	0	0
talk.politics.misc	0	0	0	0	1	6	0	11	2	9	42	24	2	114	8	0	2	0	0	0
talk.religion.misc	0	0	0	0	0	12	0	9	1	4	27	20	5	86	6	0	1	2	2	2

TF-IDF + LogReg 主要混淆对

1. talk.politics.misc 被预测为 talk.politics.guns : **94** 次
2. talk.religion.misc 被预测为 soc.religion.christian : **71** 次
3. alt.atheism 被预测为 soc.religion.christian : **64** 次
4. comp.windows.x 被预测为 comp.graphics : **58** 次
5. talk.religion.misc 被预测为 alt.atheism : **49** 次
6. comp.sys.ibm.pc.hardware 被预测为 comp.os.ms-windows.misc : **37** 次
7. comp.sys.ibm.pc.hardware 被预测为 sci.electronics : **35** 次
8. rec.autos 被预测为 rec.sport.baseball : **34** 次
9. comp.os.ms-windows.misc 被预测为 comp.sys.ibm.pc.hardware : **33** 次
10. comp.sys.mac.hardware 被预测为 comp.sys.ibm.pc.hardware : **32** 次
11. rec.sport.hockey 被预测为 rec.sport.baseball : **31** 次
12. comp.os.ms-windows.misc 被预测为 comp.graphics : **28** 次

LLM 主要混淆对

1. talk.politics.mideast 被预测为 soc.religion.christian : **167** 次
2. talk.politics.guns 被预测为 soc.religion.christian : **130** 次
3. rec.motorcycles 被预测为 rec.autos : **115** 次
4. talk.politics.misc 被预测为 soc.religion.christian : **114** 次
5. alt.atheism 被预测为 soc.religion.christian : **106** 次
6. comp.sys.ibm.pc.hardware 被预测为 rec.autos : **103** 次
7. comp.sys.mac.hardware 被预测为 rec.autos : **98** 次
8. comp.windows.x 被预测为 sci.electronics : **95** 次
9. comp.graphics 被预测为 sci.electronics : **89** 次
10. comp.os.ms-windows.misc 被预测为 sci.electronics : **89** 次
11. sci.crypt 被预测为 sci.electronics : **86** 次
12. sci.space 被预测为 sci.electronics : **86** 次

说明: LLM 存在部分输出无法映射到 20 个标准标签的情况, 评估中记为 unknown; 在按 20 类混淆矩阵展示时, 这部分样本不计入 20x20 方阵, 详细数量见 summary 文件中的 unknown_prediction_rate。