Data Folder Specification Document

Purpose: This document is a *ground-truth specification* of the `data/` directory used for an **NLP-based Sarcasm Detection and Fact Verification system**.
It is intended to be given **as-is** to a code-generation system (e.g., Perplexity) so that model code can be written **strictly based on the real datasets**, their formats, labels, sizes, and modalities.

Canonical `data/` Directory Structure (FINAL)

```
data/
├── FEVER/
│   ├── fever_train.jsonl
│   └── fever_test.jsonl
│
├── LIAR/
│   ├── README
│   ├── train_formatted.csv
│   ├── test.tsv
│   └── valid.tsv
│
├── mmsd2/
│   ├── dataset_image/
│   │   ├── <image_id>.jpg
│   │   └── ... (24,636 images)
│   └── text_json_final/
│       ├── train.json
│       ├── test.json
│       └── valid.json
│
├── MRPC/
│   ├── train.tsv
│   ├── test.tsv
│   └── dev.tsv
│
├── mustard_repo/
│   ├── data/
│   │   ├── sarcasm_data.json
│   │   ├── bert-input.txt
│   │   ├── audio_features.p
│   │   ├── split_indices.p
│   │   └── videos/
│   │       ├── utterances_final/
│   │       └── context_final/
│   │
│   ├── images/
│   │   └── utterance_example.jpg
│   │
│   ├── visual/
│   │   ├── c3d.py
│   │   ├── i3d.py
│   │   ├── dataset.py
│   │   ├── extract_features.py
│   │   ├── save_frames.sh
│   │   └── README.md
│   │
│   ├── extract_audio_features.py
│   ├── extract_audio_files.sh
│   └── README.md
│
├── paranmt/
│   ├── para-nmt-5m-processed.txt
│   └── README
│
├── quora/
│   ├── train.csv
│   └── test.csv
│
├── sarc/
│   └── train-balanced-sarcasm.csv
│
├── Sarcasm Headlines/
│   └── Sarcasm_Headlines_Dataset.json
│
├── sarcnet/
│   └── SarcNet Image-Text/
│       ├── Image/
│       │   ├── 1.jpg
│       │   ├── 2.jpg
│       │   └── ... (3,335+ images)
│       ├── SarcNetTrain.csv
│       ├── SarcNetVal.csv
│       └── SarcNetTest.csv
│
```

Dataset-Level Technical Specifications

1. FEVER — Fact Verification

Modality: Text only
Task: Claim verification (3-class classification)

Files & Sizes:

- `fever_train.jsonl` (\~61.6 MB)
- `fever_test.jsonl` (\~7.5 MB)

Schema (per JSON line):

- `id` (int)
- `claim` (string)
- `label` (string): `SUPPORTS`, `REFUTES`, `NOT ENOUGH INFO`
- `evidence_annotation_id` (int)
- `evidence_id` (int, -1 possible)
- `evidence_wiki_url` (string)
- `evidence_sentence_id` (int, -1 possible)

Primary Model Inputs:

- `claim`

Target Label:

- `label`

---

2. LIAR — Political Fact Checking

Modality: Text + metadata
Task: Fake news / truthfulness classification

Files & Sizes:

- `train_formatted.csv` (\~1.1 MB)
- `test.tsv` (\~295 KB)
- `valid.tsv` (\~295 KB)

Original TSV Columns (14 columns):

1. ID
2. Label (6-class)
3. Statement

4. Subject
5. Speaker
6. Job Title
7. State Info
8. Party Affiliation
9. Barely True Count
10. False Count
11. Half True Count
12. Mostly True Count
13. Pants on Fire Count
14. Context

Labels (Original):

- True
- Mostly True
- Half True
- Mostly False
- False
- Pants on Fire

Primary Model Inputs:

- `Statement`

Target Label:

- `Label` (optionally collapsed outside this document)

---

3. MMSD2 — Multimodal Sarcasm Detection (Text + Image)

Modality: Text + Image
Task: Binary sarcasm detection

Files & Sizes:

- Images: 24,636 files (\~2.5 GB total)
- `train.json` (\~2.3 MB)
- `test.json` (\~291 KB)
- `valid.json` (\~292 KB)

JSON Schema:

- `text` (string)
- `label` (int): 0 = non-sarcastic, 1 = sarcastic
- `imageid` (string, maps to image filename)

Primary Model Inputs:

- `text`
- Image file resolved via `imageid`

Target Label:

- `label`

---

4. MRPC — Microsoft Research Paraphrase Corpus

Modality: Text (sentence pairs)
Task: Paraphrase identification

Files & Sizes:

- `train.tsv` (\~944 KB)
- `test.tsv` (\~447 KB)
- `dev.tsv` (\~106 KB)

TSV Schema:

1. `Quality` (int): 1 = paraphrase, 0 = non-paraphrase
2. `#1 ID` (int)
3. `#2 ID` (int)
4. `#1 String` (string)
5. `#2 String` (string)

Primary Model Inputs:

- Sentence 1
- Sentence 2

Target Label:

- `Quality`

---

5. MUStARD — Multimodal Sarcasm (Text + Audio + Video)

Modality: Text + Audio + Video + Context
Task: Binary sarcasm detection

Core Statistics:

- 690 utterances

- Perfect 50/50 sarcasm balance

Key Files:

- `sarcasm_data.json`
- `bert-input.txt`
- `audio_features.p`
- `split_indices.p`
- Videos in `utterances_final/` and `context_final/`

JSON Schema (sarcasm_data.json**)**:

- `utterance` (string)
- `speaker` (string)
- `context` (array of strings)
- `context_speakers` (array of strings)
- `show` (string)
- `sarcasm` (boolean)

Primary Model Inputs:

- `utterance`
- `context`
- Audio feature vector
- Video clip

Target Label:

- `sarcasm`

---

6. ParaNMT-5M — Large-Scale Paraphrase Corpus

Modality: Text (sentence pairs)
Task: Paraphrase generation / similarity

Files & Sizes:

- `para-nmt-5m-processed.txt` (\~520 MB)

File Format (TSV per line):

1. Reference sentence (string)
2. Paraphrase sentence (string)
3. Paragram similarity score (float)

Primary Model Inputs:

- Sentence 1
- Sentence 2

Target Signal:

- Similarity score (optional usage)

---

7. Quora Question Pairs

Modality: Text (question pairs)
Task: Duplicate question detection

Files & Sizes:

- `train.csv` (\~60.5 MB)
- `test.csv` (\~455 MB)

Train CSV Schema:

- `id`
- `qid1`
- `qid2`
- `question1`
- `question2`
- `is_duplicate` (0/1)

Primary Model Inputs:

- `question1`
- `question2`

Target Label:

- `is_duplicate`

---

8. SARC — Balanced Reddit Sarcasm Corpus

Modality: Text + context
Task: Binary sarcasm detection

Files & Sizes:

- `train-balanced-sarcasm.csv` (\~249 MB, \~1.3M rows)

CSV Schema:

- `label` (0/1)
- `comment`
- `author`
- `subreddit`
- `score`
- `ups`
- `downs`
- `date`
- `created_utc`
- `parent_comment`

Primary Model Inputs:

- `comment`
- `parent_comment`

Target Label:

- `label`

---

9. Sarcasm Headlines Dataset

Modality: Text
Task: Binary sarcasm detection

Files & Sizes:

- `Sarcasm_Headlines_Dataset.json` (\~5.8 MB, 28,619 records)

JSON Schema (per line):

- `is_sarcastic` (0/1)
- `headline` (string)
- `article_link` (string)

Primary Model Input:

- `headline`

Target Label:

- `is_sarcastic`

---

10. SarcNet — Multimodal Image-Text Sarcasm (Multi-Label)

Modality: Text + Image
Task: Sarcasm detection with modality-specific labels

Files:

- `SarcNetTrain.csv`
- `SarcNetVal.csv`
- `SarcNetTest.csv`
- `Image/` (\~3,335 images)

CSV Schema:

- `Text`
- `Imagepath`
- `Textlabel` (0/1)
- `Imagelabel` (0/1)
- `Multilabel` (0/1)

Primary Model Inputs:

- `Text`
- Image resolved via `Imagepath`

Target Labels:

- `Textlabel`
- `Imagelabel`
- `Multilabel`

End of Document

This document is intentionally exhaustive and literal so it can be used as a **single source of truth** for dataset loading, preprocessing, and model design without ambiguity.