Text Normalization and TF–IDF on a Real Dataset

`twitter_f1` (Hugging Face)

Enrique Ulises Báez Gómez Tagle

August 15, 2025

Assignment

- Choose a real text dataset (paper or open-source repository: **Kaggle**, UCI, HuggingFace).
- Present the dataset:
  1. **Source**: cite paper or repository.
  2. **Domain & Topic**.
  3. **Collection Method** (if provided by the source).
  4. **Purpose**: what NLP tasks it enables.
- Apply normalization: lowercasing, expand contractions, remove punctuation & stopwords, stemming/**lemmatization**.
- Vectorize with **TF–IDF**; show **vocabulary size** and **sample vectors**.
- Prepare the pipeline: **dataset → normalization → TF–IDF → observations**.

Dataset: `twitter_f1`

Source (citation):

- Hugging Face Datasets: **Malekith/twitter_f1**. Available at `https://huggingface.co/datasets/Malekith/twitter_f1` (accessed August 15, 2025).

Domain & Topic:
Public tweets about Formula 1 (teams, drivers, qualifying, races, results, support/celebration messages).

Collection Method (not specified in the dataset card):

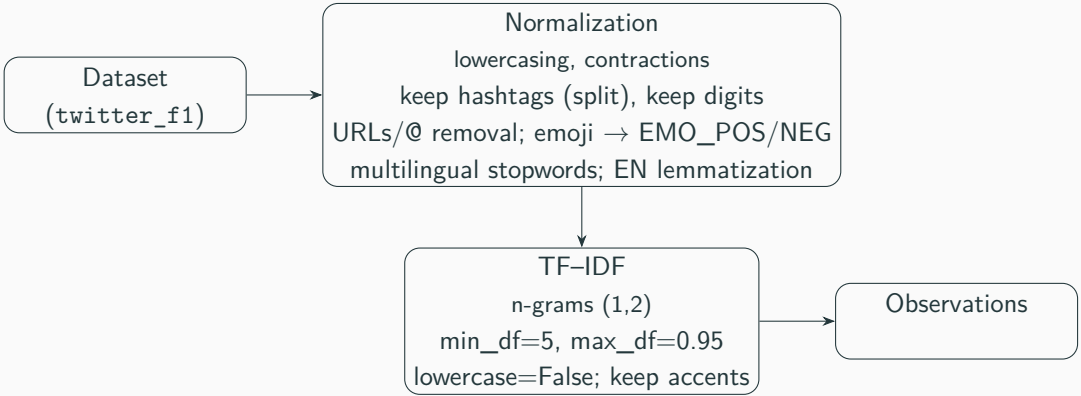
> *Based on the title and tags, this appears to consist of Twitter/X posts related to Formula 1, but no explicit collection details are provided in the dataset card.*

Languages: English mixed with other languages in the text.
License: not specified in the dataset card.

Possible NLP Uses

- Sentiment analysis, keywording.
- Information retrieval/search, clustering.
- Event/entity-focused analysis in sports media.

Workflow Overview



Normalization Steps Implemented

- **Lowercasing**: convert all text to lowercase.
- **Contractions expansion**: e.g., `it's → it is`.
- **URLs & @mentions**: removed; **hashtags preserved** (strip #, split CamelCase; special-case `GP` as `gp`).
- **Emoji tokens**: mapped to sentiment tokens (`EMOPOS`/`EMONEG`).
- **Keep alphanumerics**: preserve domain tokens like `f1`, `p1`/`p2`, `q3`; drop pure punctuation.
- **Stopwords**: multilingual set (EN + ES/IT/FR) + `{amp, rt}`.
- **Lemmatization**: spaCy English *only for EN* texts; non-EN texts skip lemmatization (cleaned tokens kept).

Vectorization (TF–IDF)

- **Vectorizer**: `TfidfVectorizer(ngram_range=(1,2), min_df=5, max_df=0.95, lowercase=False, strip_accents=None)`.
- **Matrix shape**: `(n_docs, vocab_size)`.
- **Results**:
  - `TF-IDF matrix shape: (4538, 2533)`.
  - `Vocabulary size: 2533`.
- **Why these parameters**:
  - `ngram_range=(1,2)`: captures key expressions (e.g., *pole position*, *hard work*).

Sample Vectors (sparse view)

From `../code/outputs/tfidf_sample_sparse.csv` (first 10 rows).

feature	tf-idf	doc
entire team	0.2772619348271273	0
lovely	0.26501653242254547	0
team hard	0.26501653242254547	0
clapping hand	0.26007489547975754	0
entire	0.25781819176685666	0
let push	0.24037394260799597	0
big thank	0.22729144383365152	0
yes	0.22164159836964992	0
hard work	0.20963085678321863	0
clapping	0.20712324803616872	0

Sample Vectors (dense view)

- A dense slice for the first 5 documents was exported to: `../code/outputs/tfidf_sample_dense_first5.csv`.
- **Note**: the matrix is very wide (2533 columns).

Top-k TF–IDF terms (k=10)

Doc 0	Doc 1	Doc 2
entire team (0.2773)	hand (0.2746)	party (0.3663)
team hard (0.2650)	team clap (0.2632)	popper (0.3663)
lovely (0.2650)	clap hand (0.2473)	old (0.3525)
clapping hand (0.2601)	great work (0.2473)	happy birthday (0.3330)
entire (0.2578)	today great (0.2363)	clapping hand (0.3163)
let push (0.2404)	forward race (0.2341)	birthday (0.3018)
big thank (0.2273)	work team (0.2319)	clapping (0.2519)
yes (0.2216)	ok hand (0.2184)	support (0.2293)
hard work (0.2096)	pole position (0.2156)	big (0.2280)
clapping (0.2071)	clap (0.2065)	year (0.2182)

Top Terms per Document (k=10)

- **Doc 0**: `entire team, team hard, lovely, clapping hand, entire, let push, big thank, yes, hard work, clapping`
- **Doc 1**: `hand, team clap, clap hand, great work, today great, forward race, work team, ok hand, pole position, clap`
- **Doc 2**: `party, popper, old, happy birthday, clapping hand, birthday, clapping, support, big, year`

Observations

- **Corpus size**: 4,538 documents.
- **Vocabulary (post-normalization)**: 2,533 terms (uni/bi-grams).
- **Themes**: stronger F1 markers retained (e.g., *f1, p1/p2, q3*), GP hashtags (e.g., *bahrain gp*), and emoji-derived tokens (e.g., *clapping hand*).
- **Multilingual**: EN + ES/IT/FR present.
- **Utility**: sentiment analysis, keywording, information retrieval, clustering, event analysis.

Reproducibility & Artifacts

- Script: `normalization_tf-idf.py`.
- Exported artifacts:
  - `../code/outputs/tfidf_sample_sparse.csv`
  - `../code/outputs/tfidf_sample_dense_first5.csv`
  - `../code/outputs/vocabulary.csv`
  - `../code/outputs/clean_texts.csv`

Before vs After

Original tweet
RT @MercedesAMGF1 What a quali! P1 for Lewis at #BahrainGP [trophy] [car] Let's push tomorrow!! [trophy] [car]

After normalization
what quali p1 lewis bahrain gp EMOPOS emopos let push tomorrow EMOPOS

References

- **Dataset**: Malekith. *twitter_f1*. Hugging Face Datasets. `https://huggingface.co/datasets/Malekith/twitter_f1` (accessed August 15, 2025).