EMOTION CLASSIFICATION IN ARABIC TWEETS

PROJECT OVERVIEW

This assignment focuses on classifying emotions in Arabic tweets using various machine learning and deep learning approaches. The dataset contains tweets labeled with 8 emotion categories: anger, fear, joy, love, sadness, surprise, sympathy, and none.

1- Data preprocessing (cleaning, normalization, tokenization)
2- Feature extraction (BoW, TF-IDF, Word2Vec averages)
3- Model training with traditional ML algorithms and neural networks
4- Performance evaluation and model comparison

1. Data Preprocessing

- Loaded and cleaned the dataset (10,065 tweets)
- Removed "none" labeled tweets (1,550 instances)
- Performed Arabic text cleaning:
  - Removed non-Arabic characters and diacritics
  - Normalized text (character unification)
  - Removed stopwords
  - Tokenized text
- Analyzed token frequency distribution

2. Feature Extraction

Implemented three different feature extraction methods:

- Bag-of-Words (BoW)
- TF-IDF
- Word2Vec Embeddings (using pre-trained Arabic vectors)

3. Modeling Techniques

A. Traditional ML Models (with GridSearchCV tuning):

1. Naive Bayes (Multinomial for BoW/TF-IDF; Gaussian for Word2Vec)
2. SVM (RBF kernel tuned for C and gamma)
3. Decision Tree (tuned for max_depth and min_samples_split)
4. Random Forest (tuned for n_estimators and max_depth)
5. AdaBoost (tuned for n_estimators and learning_rate)

B. Neural Networks:

1. Dense Network:
    - Architecture: Input → Dense(64) → Dropout(0.3) → Dense(32) → Output
    - Applied to TF-IDF and Word2Vec features.
2. Bidirectional LSTM:
    - Embedding layer initialized with Word2Vec

EVALUATION METRICS

Model	Feature	Accuracy	F1	Best params
random forest	TF-IDF	62%	62%	n_estimators=200, max_depth=None
SVM	Word2Vec	67%	67%	C=1, gamma='auto', kernel='rbf'
LSTM	Word2Vec	66.5%	-	Embedding + LSTM(128)
AdaBoost	BOW	36.3%	33%	n_estimators=100, learning_rate=1.0
Dense NN	Word2Vec	66%	-	Dense(64) → Dropout(0.3) → Dense(32) → Dropout(0.3)
Dense NN	TF-IDF	63%	-	same

CONCLUSION

BEST MODEL: SVM WITH WORD2VEC FEATURES

- ACCURACY: 67.06%
- MACRO F1-SCORE: 67.19%
- WHY: OUTPERFORMED ALL OTHER MODELS IN ACCURACY AND F1 WHILE BEING COMPUTATIONALLY EFFICIENT.