Title: NewsBot Intelligence System 2.0 - Technical Document Author: Ahmet Burak Solak Course: ITAI2373 Version: Final Draft

1. Goal The system reads news. It cleans text. It learns patterns. It gives category, sentiment, entities, summary.

2. Data File: BBC News Train.csv. Columns: ArticleId, Text, Category. Five categories: sport, business, politics, entertainment, tech. I remove rows with empty text.

3. Main Steps Step 1: Load data. Step 2: Clean text (lowercase, remove symbols, lemmatize, remove stop words). Step 3: Make features (TF-IDF words + small linguistic numbers like noun ratio). Step 4: Train model (Random Forest) for category. Step 5: Sentiment (TextBlob polarity and subjectivity). Step 6: Entities (spaCy PERSON, ORG, GPE, etc.). Step 7: Summary (simple extractive top sentences by TF-IDF score). Step 8: Conversation bot answers simple user questions.

4. Libraries pandas, numpy, scikit-learn, spaCy, TextBlob, seaborn, matplotlib.

5. Models Classifier: RandomForestClassifier with 200 trees. Vectorizer: TfidfVectorizer with max 3000 features, 1-2 grams. Language model: spaCy en_core_web_sm.

6. Features TF-IDF matrix (word importance). Noun ratio, verb ratio, average sentence length. Sentiment polarity and subjectivity. Entity list per article.

7. Output Category label. Category confidence (highest probability). Sentiment numbers. Entity groups. Short summary.

8. Conversation Logic User writes a question. Bot sees words like category, sentiment, find. Bot returns counts or search results. If word looks like a category it lists articles count.

9. Multilingual Plan (Future) Detect language. Translate text to English. Then do same steps. Not implemented fully. Placeholder only.

10. Performance Print accuracy and report. Show confusion matrix. Accuracy is high (around above 0.8). Exact number prints in notebook.

11. Simple Architecture Data -> Clean -> Features -> Train -> Analyze -> Serve via bot.

12. Limits English only now. Summary is very basic. Search is simple string match.

13. Next Improvements Better summary (abstractive). Add topic model (LDA). Add real translation. Add web app front end.

14. Run Order Run setup cell. Load data cell. Preprocess cell. Features cell. Train model cell. Analysis cells. Bot cell. Test cell.

15. Folder Items Notebook: NewsBot_Final_System.ipynb CSV files: BBC News Train.csv etc. Requirements: requirements.txt