Natural Language Processing

Formality Checker for Text Analyzing Text for Formal and Informal Tone Using NLP

Report Synopsis

Submitted By:

22-CS-110 (Muhammad Sohaib)

Submitted By:

Dr. Zahid Mehmood

Department of Computer Science University of Engineering and Technology, Taxila

Fall 2025

Chapter 1: Introduction

1.1 Background

Textual communication is now among the most normal modes of interaction on personal, educational, and business levels since it is extremely quick, convenient, and highly accessible to most people. Nonetheless, digital communication has also expanded at a very high rate, which has posed challenges in ensuring the appropriate tone of writing on various platforms. Individuals tend to alternate informal messages in social media and professional writing in emails or reports giving rise to a variation in tone and the unintended use of informal phrases in formal contexts. This discrepancy may have an adverse influence on clarity, professionalism, and the general impact of the message. Natural Language Processing (NLP) and Machine Learning (ML) become inevitable with the rise in the use of digital communication. These technologies allow analyzing the language style of a text and deciding whether it is formal or informal. Through vocabulary, structure of grammar and writing form, such systems would be able to automatically assist users in ensuring that they use the right tone in their communication.

1.2 Problem Statement

The volume of digital writing that is being generated in the contemporary communication systems is expanding exponentially and this is diminishing the clarity of written communication and posing challenges to people and organizations to retain an appropriate tone. Old methods of tone-checking, including manual proofreading, or even simple rule-based systems are reflective and can easily be fooled by the subtle informal aspects of language such as contraction, slang or colloquial phrasing. Consequently, there is great demand in a smart, automatic technology which will know the structure and style of written language and categorize it correctly to save the users the embarrassment of using the wrong tone inadvertently and to have the words communicate clearly and steadily.

1.3 Aim & Objectives

Aim:
To create a powerful Text Formality Detection system based on the NLP techniques of linguistic analysis and the tone classification of the machine learning method.

Objectives:

1. To preprocess raw text data (eliminate unwanted elements, such as stop words, punctuation etc.) and standardize language (stemming/lemmatization).

2. To find significant linguistic data in the text with the help of such methods as TF-IDF (Term Frequency Inverse Document Frequency) or Bag of Words.

3. To train and compare the performance of various classification algorithms (e.g., Naïve Bayes, Support VectorMachine or Random Forest).

4. To evaluate the performance using performance measures on the model.**Relevant**

Data Set Detail

Name: formal_informal_dataset

Description: Aim:

The dataset is very popular in the field of formality detection because it has a big set of texts which are characterized by the labels of formal and informal. These are all gathered online in various places like the forums, social sites and paraphrased versions that are done by professional editors. The text samples are already categorized into such categories as formal, informal and neutral, which makes them appropriate to train and evaluate the classification models.

Performance Evaluation Matrix

To measure the effectiveness of the model, I will theoretically apply the following confusion matrix-derived metrics:

- **Accuracy:** Accuracy measures the overall correctness of the model. It is the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- **TP (True Positives):** Text correctly identified as formal
- **TN (True Negatives):** Text correctly identified as informal
- **FP (False Positives):** Informal text incorrectly marked as formal
- **FN (False Negatives):** Formal text incorrectly marked as informal
- **Precision:** Precision indicates how many of the texts predicted as formal were formal. High precision ensures that the model avoids misclassifying informal text as formal.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** Recall measures the ability of the model to capture all actual formal texts. It calculates how many actual formal texts were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** F1-score is the harmonic means of precision and recall and provides a balance between the two metrics. It is especially useful when there is an imbalance between formal and informal text samples.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

1.4 Applications

• **Email and Messaging Platforms:** The system can be installed into email services such as Gmail, Outlook or available messaging applications and automatically scan the tone of the incoming messages and mark informal or unprofessional messages.

• **Enterprise Communication:** This system would help organizations to maintain professional standards and formal echo in internal and external email, reports and memos to ensure that they communicate in a professional manner.

• **Customer Support:** The system can study customer support responses or tickets and determine the formal and informal responses to retain the same professional communication.

Chapter 2:

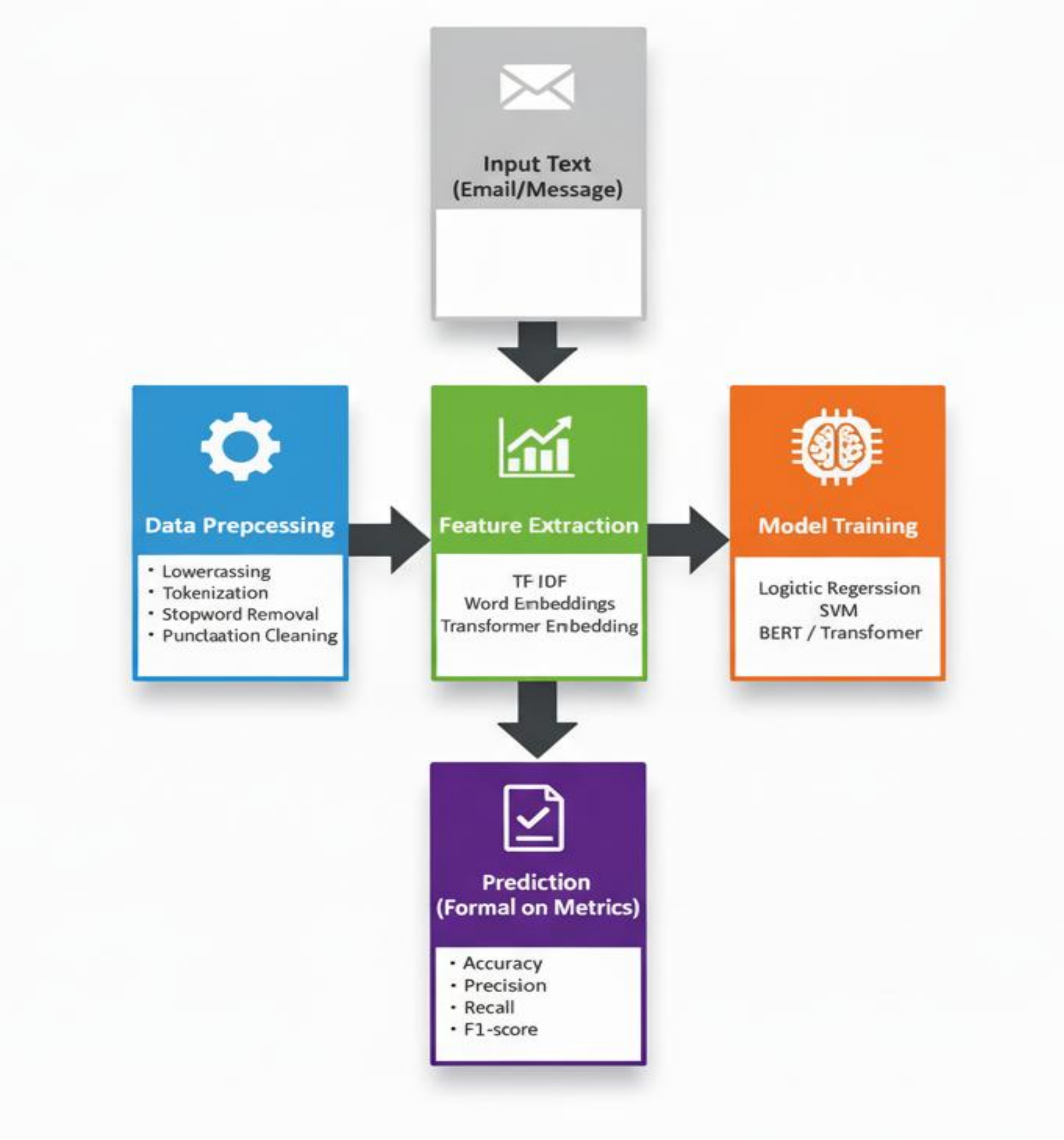
2.1 Related work

Author	**Model Used**	**Performance/ Accuracy**	**Dataset Used**	**Key Limitations**
Daryna Dementiea	Statistical baselines, Char BiLSTM, Transformers	Accuracy / F1: ~79–85% depending on model	GYAFC, X-FORMAL	Limited to binary classification domain-specific
Kunal Chawla	BART seq-to-seq + language model discriminator + mutual information loss	BLEU human eval: improved formality & diversity	GYAFC + Yelp / Amazon	Requires parallel data outputs may be repetitive BLEU may not fully capture style/meaning trade-offs
Yi Zhang	Transformer seq-to-seq + data augmentation	State-of-the-art BLEU on GYAFC with augmented data	GYAFC + GEC corpora	Augmentation may generate noisy pairs; synthetic data may skew style distribution relies on quality of back-translation
Zichao Yang	Generator + target-domain language model as discriminator	NLL reduction; competitive unsupervised performance	Non-parallel corpora	Style/content disentanglement implies no parallel data
Parastoo Falakaflaki	Fa-BERT2BERT:	BLEU, BERT	Persian	Low-resource

	BERT-to-BERT with consistency learning	Score, ROUGE-L improved style metrics	informal/formal corpus	language limited parallel data model complexity & training cost high

Chapter 3

3.1 Methodology & Block Diagram



Input Text
(Email/Message)
Data Prepcessing
• Lowercassing
• Tokenization
• Stopword Removal
• Punctaation Cleaning
Feature Extraction
TF IDF
Word Enbeddings
Transformer Enbedding
Model Training
Logictic Regerssion
SVM
BERT / Transfomer
Prediction
(Formal on Metrics)
• Accuracy
• Precision
• Recall
• F1-score