# INTRODUCTION TO
# NATURAL LANGUAGE PROCESSING

CLASS NOTES

Mahanth Yalla
M. Tech-AI, IISc

# Preface

These notes are based on the lectures delivered by **Prof. Danish Pruthi** in the course **DS 207 - Introduction to Natural Language Processing** at Indian Institute of Science (IISc) Bengaluru - Jan Semester 2025. The notes are intended to be a concise summary of the lectures and are not meant to be a replacement for the lectures. The notes are written in LaTeX .

# Disclaimer

These notes are not official and may contain errors. Please refer to the official course material for accurate information.

*"I cannot guarantee the correctness of these notes. Please use them at your own risk".*

- Mahanth Yalla

## Contribution

If you find any errors or have any suggestions, please feel free to open an issue or a pull request on this GitHub repository, I will be happy to incorporate them.

## Feedback

If you have any feedback or suggestions on the notes, please feel free to reach out to me via social media or through mail mahanthyalla [at] {iisc [dot] ac [dot] in , gmail [dot] com}.

**Acknowledgements:** I would like to thank Prof. Danish Pruthi for delivering the lectures and providing the course material. I would also like to thank the TAs for their help and support.

# Contents

# Chapter 1

# Text Classification

## 1.1 Introduction

### 1.1.1 Introduction 2

- Text Classification is a supervised learning problem.
- It is a type of document classification where the goal is to categorize documents into a fixed set of categories or classes.
- It is a fundamental problem in NLP.
- It is used in various applications like spam filtering, sentiment analysis, language identification, genre classification, etc.

### 1.1.2 Applications

- Spam Filtering: Classify emails as spam or not spam.
- Sentiment Analysis: Classify documents as positive, negative or neutral.
- Language Identification: Classify documents into different languages.
- Genre Classification: Classify documents into different genres like news, sports, politics, etc.

### 1.1.3 Challenges

- High Dimensionality: Text data is high dimensional.
- Data Sparsity: Text data is sparse.
- Synonymy: Different words can have the same meaning.
- Polysemy: Same word can have different meanings.
- Ambiguity: Text data can be ambiguous.
- Overfitting: Overfitting is a common problem in text classification.

### 1.1.4 Approaches

**Naive Bayes**

Naive Bayes is a simple probabilistic classifier based on Bayes theorem. sumes that the features are independent given the class label. It is widely used in text classification.

# Chapter 2

# Word Representations

## 2.1 Introduction

- Words are the basic building blocks of any language.
- Words are the smallest unit of meaning
- Words are represented in the form of vectors in NLP.
- Word representations are used in various NLP tasks like text classification, machine translation, sentiment analysis, etc.