**Text

Description automatically generated**

**Word Embeddings**

**Coursework 2024/2025**Steve Galea (0015902H)

Study-unit: **Topics in Applied Data Science**

Code: **CIS5231**

Lecturer: Prof John Abela

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

Declaration Plagiarism is defined as “the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

I, the undersigned, declare that the assignment submitted is my work, except where acknowledged and referenced.

I understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

A close up of a signature

AI-generated content may be incorrect.Work submitted without this signed declaration will not be corrected, and will be given zero marks.

STEVE GALEA \_\_\_\_\_\_\_\_\_\_

Student Name Signature

Word Embeddings Coursework 2024/2025

CIS5231

Course Code Title of work submitted

15/06/2025

Date

# Table of Contents

# Introduction

# Literature Review Of Word Embedding Techniques

# Detailed Explanation of the:

## Acquisition and Pre-Processing

The News Articles[[1]](#footnote-1) dataset was acquired from Kaggle. It contains ~2500 scraped news articles and headlines related to business and sports from 2015 to 2017. For this assignment, since computations will be carried out on a CPU, the business articles will be selectively focused on.

Pre-Processing Steps:

1. Text Normalisation: Convert all text to lowercase to maintain consistency.

2. Punctuation Removal: Remove punctuation using regular expressions to clean the text.

3. Tokenisation: Split text into individual tokens (words).

4. Stop-Word Removal: Filter out common stop words to reduce noise and focus on semantically rich words.

5. Vocabulary Limiting: Construct a vocabulary of the top 10,000 most frequent words, replacing infrequent words with an <UNK> token.

6. Indexing: Convert tokens into numerical indices based on the constructed vocabulary

## Model

## Training Process

## Visualisation Techniques Used

# Discussion of results

# Challenges Encountered

# Possible Enhancements

# Conclusion

# ChatGPT usage

Chat GPT was used similar

A screenshot of a computer program

AI-generated content may be incorrect.

1. https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles [↑](#footnote-ref-1)