Table of Contents

[**Experiment 1: Text Classification using Word2Vec and various classification Models 2**](#_30j0zll)

[**Experiment 2: Evaluation of model (Word2Vec and classification Algorithms) on new data 4**](#_1fob9te)

[**Experiment 3: Text Classification Using TF-IDF and various Classification Models 5**](#_3znysh7)

[**Experiment 4: Evaluation of model (TF-TDF and classification Algorithms) one new data 6**](#_2et92p0)

[**Experiment 5: Feature Engineering and it’s correlation. 7**](#_gk9nmwvoh8bv)

# **Experiment 1: Text Classification using Word2Vec and various classification Models**

1. **Overview:** This report summarises the process and results of a text classification project. The workflow includes the following steps:
   1. **Data preprocessing**
   2. **Word2Vec embedding generation**
   3. **Feature Normalization**
   4. **Training multiple classification algorithms**
2. **Dataset:** 
   1. **Training Set:**  Around 10,000 samples with text and label columns.
   2. **Test Set:** Around 2,500 samples with text and label columns.
3. **Models Used:**
   1. **XGBoost Classifier**
      1. Evaluation metric: logloss
      2. Random state: 47
   2. **Support Vector Machine (SVM)**
   3. **Naive Bayes (MultinomialNB)**
   4. **Random Forest Classifier**
      1. Random state: 47
4. **Embedding Model: Word2Vec**

The Word2Vec model was trained with the following parameters:

* 1. **Vector size:** 100 (dimension of the word embeddings).
  2. **window:** 5(context window size around a word).
  3. **Min count:** 2 (ignores words appearing fewer than 2 times).
  4. **workers:** 4 (number of parallel threads used for training).
  5. **sg:** 0 (uses CBOW instead of Skip-gram).

1. **Normalisation:**
   1. Applied Min-Max scaling to normalize feature vectors into the range [0,1].
2. **Results**

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 Score(%) |
| --- | --- | --- | --- | --- |
| XGBoost | 95.71 | 94.46 | 97.06 | 95.74 |
| SVM | 96.05 | 95.64 | 96.46 | 96.05 |
| Naive Bayes | 72.10 | 67.71 | 83.88 | 74.93 |
| Random Forest | 94.51 | 92.36 | 96.98 | 94.61 |

With extracted Features

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 Score(%) |
| --- | --- | --- | --- | --- |
| XGBoost | 97.34 | 96.36 | 98.36 | 97.35 |
| SVM | 93.87 | 92.62 | 95.26 | 93.92 |
| Naive Bayes | 80.24 | 82.69 | 76.20 | 79.31 |
| Random Forest | 96.19 | 95.57 | 96.81 | 96.18 |

# **Experiment 2: Evaluation of model (Word2Vec and classification Algorithms) on new data**

1. **Load and Balance the Dataset:**
   1. A balanced dataset was created with train set:
      1. **2,000 samples** of human-generated text.
      2. **2,000 samples** of AI-generated text.
2. **Model Evaluation:**

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 Score(%) |
| --- | --- | --- | --- | --- |
| XGBoost | 65.48 | 66.11 | 63.50 | 64.78 |
| SVM | 65.15 | 64.46 | 67.55 | 65.97 |
| Naive\_Bayes | 61.25 | 58.09 | 80.80 | 67.59 |
| Random Forest | 65.75 | 64.11 | 71.55 | 67.63 |

# **Experiment 3: Text Classification Using TF-IDF and various Classification Models**

1. **Overview:** This experiment focuses on classifying text using **TF-IDF embeddings** as feature vectors and multiple machine learning algorithms for classification.
2. **Dataset:** 
   1. **Training Set:**  Around 10,000 samples with text and label columns.
   2. **Test Set:** Around 2,500 samples with text and label columns.
3. **Models Used:**
   1. **XGBoost Classifier**
      1. Evaluation metric: logloss
      2. Random state: 47
   2. **Support Vector Machine (SVM)**
   3. **Naive Bayes (GaussianNB)**
   4. **Random Forest Classifier**
      1. Random state: 47
4. **TF-IDF Vectorizer (with maximum features: 25,000)**
5. **Results**

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 Score(%) |
| --- | --- | --- | --- | --- |
| XGBoost | 97.30 | 97.08 | 97.24 | 97.16 |
| SVM | 96.11 | 95.98 | 95.82 | 95.90 |
| Naive Bayes | 76.67 | 84.96 | 61.82 | 71.57 |
| Random Forest | 95.20 | 95.67 | 94.15 | 94.90 |

# 

# **Experiment 4: Evaluation of model (TF-TDF and classification Algorithms) one new data**

1. **Load and Balance the Dataset:**
   1. A balanced dataset was created with train set:
      1. **2,000 samples** of human-generated text.
      2. **2,000 samples** of AI-generated text.
2. **Model Evaluation Results:**

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 Score(%) |
| --- | --- | --- | --- | --- |
| XGBoost | 60.80 | 61.06 | 60.80 | 60.57 |
| SVM | 62.98 | 62.98 | 62.98 | 62.97 |
| Naive Bayes | 56.50 | 58.08 | 56.50 | 54.27 |
| Random Forest | 58.20 | 58.57 | 58.20 | 57.75 |

# 

# **Experiment 5: Feature Engineering and it’s correlation.**

