LAB2 Homework Report

Introduction

> In this Homework, the major difficulty I face is not on model building, but the time running model is to slow. I oftenly face a situation when I left the computer a long, the running section shut down by hibernate mode.
> Eventually I build a Random forest model with 0.31 score. I also plan to try other model but my schedule make me compromise.
> In this model I build a TFIDF with randomforest classfier.

Preprocessing

```
def preprocess_text(text):
    # Handle missing or non-string entries
    if not isinstance(text, str):
        return ""

    # Convert to lowercase
    text = text.lower()

    # Remove URLs (http://, https://, www)
    text = re.sub(r'http\S+|www\S+|https\S+', '', text)

    # Remove mentions (@username) and hashtags (#hashtag)
    text = re.sub(r'@\w+|#\w+', '', text)

    # Remove special characters, punctuation, and numbers
    text = re.sub(r'[^a-zA-Z\s]', '', text)

    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]

    # Lemmatize tokens
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in tokens]

    # Rejoin tokens back into a single string
    return ' '.join(tokens)

# Apply preprocessing to the 'text' column
train_tweet['cleaned_text'] = train_tweet['text'].apply(preprocess_text)
```

```
# Check the results
print(tweets[['text', 'cleaned_text']].head())
```

In preprocessing step, I remove URLS,mention,special character and stopwords,convert data to lower case and tokenize them.

Model

My model first using TFIDF building new feature with maximum of 1500. Then run a Randomforest classifier model.