AIC-201 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-201:00020
Supervised Learning for Structured Data

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

1. Feature Selection
    a) Data Volume: Get a sufficiently large dataset to ensure reliable statistical analysis and meaningful insights.
    b) Language Filtering: Apply language detection to select tweets in languages like English, Thai or others.
    c) Geolocation Filtering: Focus on tweets that have geolocation data from Thailand or contain relevant keywords tied to Thai tourist destinations.
    d) Timeframe: Collect tweets from both peak tourist seasons of off-peak periods to capture various sentiments or views of tourists.
    e) Content Filtering: Remove spam, retweets, advertisements and content that are not related to the analysis.

2. Data Preprocessing
    a) Tokenization: Splitting messages into individual words, thus, making it easier for analysis
    b) Lowercasing: Avoid distinguishing between words with or without uppercase letters.
    c) Removal of Punctuation and Special Characters
    d) Hashtags: Convert hashtags into keywords
    e) URL and Emojis: Remove URLs and translate emojis into textual form to see what the emoji contributes to the tweets' view
    f) Stopwords: Remove meaningless words such as is, and, the, and other words that do not give useful information for sentiment analysis
    g) Slangs: Expand abbreviations such as lol and convert slangs into formal language to improving model understanding.

3. Feature Engineering
    a) Features for Training the models:
        a) Bag of words – Frequency of words in dataset
        b) TF-IDF (Term Frequency-Inverse Document Frequency): Weighing terms on their frequency
        c) N-grams: Multi-word combinations
        d) Part-of-speech Tagging: Identifying adjectives, adverbs and others – key for sentiment indication
        e) Emoji Encoding: Translate emoji symbols to sentiment scores to capture emotional context of tweets
    b) Advantages and Limitations
        a) Bag of words and TF-IDF – Basic word frequency analysis but fail to capture context
        b) N-grams – Useful for adding context but can increase feature complexity

4. Attached in Canvas

AIC-201 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-201:00020
Supervised Learning for Structured Data

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

5. Model Selection – BERT (Bidirectional Encoder Representations from Transformers) or similar ones. The reasons behind the selection are:
   a) It is able to understand full context of words
   b) Pre-trained on massive text corpora like Wikipedia, make it highly proficient in language understanding
   c) It is also good at interpreting short or informal texts

6. Handling Imbalanced Data
   a) Use oversampling to increase representation of Positive and Negative tweets or sample-less neutral tweets.
   b) Assign higher weights to the Positive and Negative class during training so that it prioritizes these categories.

7. Evaluation Metrics
   a) Use metrics such as Precision, Recall and F1 score to evaluate how well the model is distinguishing between Positive, Negative and Neutral sentiments.
   b) Use the confusion matrix to show where the model is making errors
   c) Accuracy can also be used but may not be sufficient if used alone

8. Error Analysis
   a) The model might confuse mild negative sentiments with neutral ones which might impact sentiment accuracy, which, can be solved by training with more examples of mild negative sentiments or using models that more nuanced datasets.

      Incorporating ensemble models could help balance misclassifications by combining multiple models and improving robustness.

9. Challenges with Unstructured Data
   a) It will be difficult to do noise handling as social media contains various noise data such as slang, typos and informal expressions

      Solution: We can use NLP library like spaCy, and apply emoji mapping tools to clean up text data before analysis.
   b) There can be several challenges when it comes to identifying sarcasm but is common tweets.

      Solution: Fine-tune transformer models with sarcasm-heavy datasets to better capture sarcasm.
   c) Short texts, which is the majority of tweets, provide limited context for accurate sentiment detection.

      Solution: Use BERT embeddings which can be used to handle short texts by considering the surrounding words in context.

AIC-201 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-201:00020
Supervised Learning for Structured Data

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

10. Further Investigation
    a) Track sentiment shifts over time to observe seasonal or event-driven changes in public opinion
    b) Identify main topics related with Positive, Negative or Neutral sentiments to understand mainstream themes better.
    c) Assess whether tweets from accounts with more followers such as influencers having a different impact on sentiment compared to those from regular users, which could provide insights into the wider social influence.