Assignment 4.1: Semantic and Sentiment Analysis

Patricia Enrique

Department of Engineering, University of San Diego

AAI-520: Natural Language Processing

Prof. Mokhtari

October 2, 2023

Process

The dataset used for the semantic and sentiment analysis is retrieved from OSF and consists of 1076 legal sentences (Ratnayaka, 2020). Each sentence is labelled corresponding to the sentiment with a label of 0 corresponding with a negative sentiment, 1 with neutral, and 2 with a positive sentiment. The dataset comes separated into training and testing with 576 and 500 sentences respectively with no null values. The distribution of sentiment in the training data is 282:172:122 for negative, neutral, and positive sentiments respectively. For this analysis, the legal sentences are tokenized, the stop words and punctuation are removed, and all words are lowercased. A random sample of the training dataset is shown in Table 1. The term frequency–inverse document frequency (TF-IDF) features are found using TF-IDF vectorization and analyzed to extract the keywords from each sentence for the semantic analysis. The sentiment analysis is completed by creating three models using three unique classifiers (Scikit Learn, 2023). The classifiers are then optimized and evaluated on accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC).

Table 1

Preprocessed legal sentences

ID	Text	Label
25	subjected humiliating sexual comments insulted...	0
365	possessing firearm convicted felon violation 1...	0
57	engaged attempted perpetration armed robbery	0
242	alleged deficiencies governments	1
478	sighted outside dominican territorial waters d...	1

Analysis

Semantic Analysis

Semantic analysis is the process of identifying the meaning of text by interpreting language structure and grammatical formats, thus recognizing relationships between words. Extracting key words from the phrases can help to gain insights into the main topics and themes discussed. This can be achieved by using TF-IDF which quantifies the importance of a word with respect to the other words. A word with a high TF-IDF score indicates an important word (Vajjala et al., 2020). For the training dataset, the TF-IDF scores are calculated and then sorted in descending order to easily identify the keywords. Table 2 summarizes the top three keywords extracted from the first three phrases in the training set and their corresponding TF-IDF score.

Table 2

Keyword extraction and TF-IDF scores

Phrase 1		Phrase 2		Phrase 3	
Keyword	TF-IDF score	Keyword	TF-IDF score	Keyword	TF-IDF score
surplusage	0.745	existing	0.450	grooming	0.328
getting	0.667	bivens	0.450	corrections	0.328
		narrow	0.450	department	0.328

A similar process can be followed to extract the key phrases with a desired number of words. For example, Table 3 contains the top three bigrams in the first three phrases from the training set and their corresponding TF-IDF score. This modification allows for more context into the keywords that are extracted and resolves some ambiguity that arises when analyzing a singular word.

Table 3

Bigram extraction and TF-IDF scores

Phrase 1		Phrase 2		Phrase 3	
Keyword	TF-IDF score	Keyword	TF-IDF score	Keyword	TF-IDF score
getting surplusage	1.0	existing scope	0.447	grooming policy	0.333
		bivens existing	0.447	corrections grooming	0.333
		narrow bivens	0.447	department corrections	0.333

Sentiment Analysis

Sentiment analysis is the process of identifying the polarity of text such as positive, negative, or neutral. Three classifiers are selected to identify the sentiment of the phrases in the testing dataset and their parameters are tuned to optimize the performance. The parameter tuning for the support vector machine (SVM), Naïve Bayes classifier, and the gradient boosting classifier affects the precision metric the most. The models are trained using the TF-IDF vectorization for both unigram and bigrams. Table 4 compares the results from the classification using unigram and bigram models after completing optimization.

Table 4

Classifier performance for sentiment analysis

Performance Metric	Classifier					
	SVM		Naïve Bayes		Gradient Boosting	
	Unigram	Bigram	Unigram	Bigram	Unigram	Bigram
Accuracy	0.456	0.452	0.454	0.456	0.462	0.454
Precision	0.748	0.762	0.487	0.637	0.454	0.572
Recall	0.456	0.452	0.454	0.456	0.462	0.454
F1 Score	0.320	0.310	0.339	0.323	0.407	0.335
AUC-ROC Score	0.626	0.546	0.615	0.578	0.590	0.588

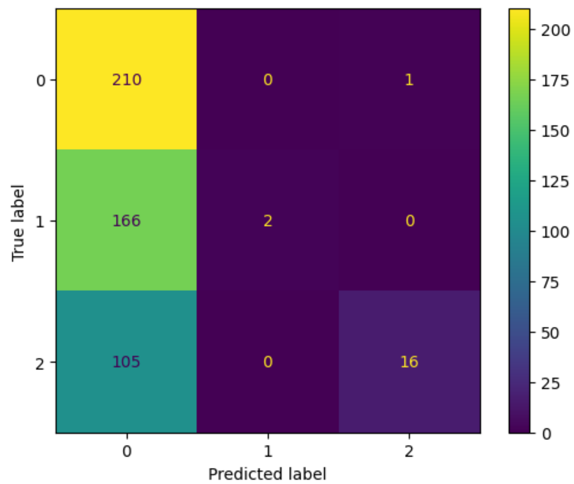
Results

The TF-IDF scores presented in Table 2 and Table 3 suggest that the three most important words and set of words in each phrase are of equal importance. This can be in part due to the fact that the phrases are short and contain few words resulting in each word being important to the context. For better results, longer phrases can be used, or phrases can be combined.

From Table 4 it can be seen that the performance of the models remains fairly consistent regardless of the classifier used or whether a unigram or bigram is used for the TF-IDF vectorization. The biggest change can be seen in the precision metric with the SVM and bigram combination achieving the best result and the gradient boosting classifier with unigram achieving the worst result. The SVM unigram model is able to correctly identify a sentiment 75% of the time as indicated by the precision and illustrated in the confusion matrix in Figure 1.

Figure 1

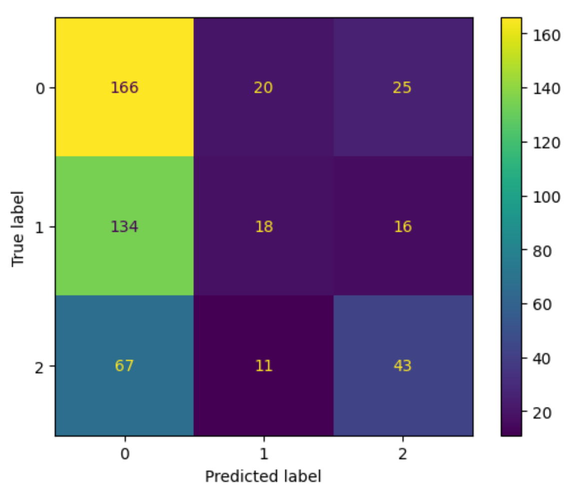
Confusion matrix resulting from the SVM model



From Figure 1 it can be seen that the model almost exclusively predicts a label of 0 corresponding with a negative sentiment and supporting the performance metrics shown in Table 4. The confusion matrix for the Naïve Bayes classifier produces similar results and the matrix for the gradient boosting classifier shows the model making more varied predictions; however still heavily predicting a negative sentiment. The gradient boosting classifier confusion matrix is shown in Figure 2.

Figure 2

Confusion matrix resulting from the gradient boosting classifier model



The poor classification performance can be attributed to the complexity of text classification and to the dataset used. Text poses various limitations to analysis as language and context such as sarcasm, irony, slang, emojis, and cultural references; and subjectivity and ambiguity such as in the cases of vague or unclear sentiments need to be considered. The data set used for this analysis is not ideal as the training set is relatively small and the label distributions are not even. Almost 50% of the phrases in the training data set have a negative sentiment which explains the tendency for models trained on this data to consistently identify the test phrases as containing a negative sentiment over a neutral or positive sentiment.

Discussion

Semantic and sentiment analysis can be utilized in various applications including customer feedback. Semantic analysis allows for quick and efficient identification of the reasons for a positive or negative review on a product or service. Instead of manually reviewing the feedback provided by customers, businesses are able to extract the keywords and phrases and identify topics or areas that are being brought up repeatedly by customers. Sentiment analysis allows businesses to understand customer preferences and opinions by measuring customer satisfaction and identifying problem areas. The sentiment of reviews can be used to recognize if changes need to be made to a business' product or service.

A specific example of semantic and sentiment analysis in customer feedback is when a business proposes an improvement to an existing product such as an improvement to an app. Users of the app can test out the improvement and provide a review. The business applies sentiment analysis to gauge if the users enjoy the improvements or if they are unhappy with the updates made. Then semantic analysis is used to extract the reasons why customers enjoy or dislike the updates and any suggestions they make. From here changes can be made to the product to improve customer satisfaction.

Conclusion

Semantic analysis is the process of identifying the meaning of text by interpreting language structure and grammatical formats. Extracting key words from the phrases using TF-IDF scores can help to gain insights into the main topics and themes discussed. Sentiment analysis is the process of identifying the polarity of text. Classifiers such as SVM, Naïve Bayes, and gradient boosting can be applied to text to predict the sentiment. Large data sets are needed to obtain good performance from the classification models. Limitations in handling complex language structures and context exist due to the complexity of the English language. Customer feedback is an application that can benefit from semantic and sentiment analysis to understand customer preferences and improve customer satisfaction.

References

Ratnayaka, G. (2020, October 25). *Legal Sentiment Analysis* [Data set]. OSF. https://osf.io/zwhm8/

Scikit Learn. (2023, September 23). sklearn.feature_extraction.text.TfidfVectorizer https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical natural language processing: A comprehensive guide to building real-world NLP systems. O'Reilly Media.