Kimya Buckner
Ling 406
Spring 2020

Sentiment Analysis Questions

1) In building the baseline system, what shallow text features make sense for the task as a first-stab approach? (a language modeling approach is a standard way to tackle this problem using, for example, a bag-of-words / a unigram model approach);

A logical first stab approach to this sentiment analysis problem is the Bag-of-Words model. This model consists of creating a lexicon of relevant words found in the documents and scoring those words in some numeric way. This is a relatively simple and straightforward approach that has historically been successfully used in sentiment analysis. Some shallow text features that work well with the bag-of-words approach are treating each individual word as a token, as well as considering only certain parts of speech (such as adjectives and adverbs) in the BoW lexicon/vocabulary.

2) What machine learning models are suitable for conducting sentiment analysis? (i.e., compare 3-4 learning algorithms of your choice). Which performs best given your features?

Naive Bayes, Logistic Regression, and Support Vector Machine Classifier are all suitable models for conducting sentiment analysis. Naive Bayes tends to perform well in document classification because it assumes that the presence of an individual feature in a document is independent of other features. This works well with a Bag-of-Words approach because context is not considered in this model. Support Vector Machine (SVM) classification also historically has

high accuracy in sentiment analysis. SVM classifiers determine the best decision boundary between vectors that belong to a given group or category and vectors that do not belong to it. As the goal of sentiment analysis is to determine the category to which a given vector belongs, SVM is a logical classification model to consider for this task. Finally Logistic Regression classifiers work best with binary or dichotomous variables such as the ones in the feature sets used here. The Support Vector Machine classifier produced the best results. As previously stated, SVM classifiers determine the best decision boundary between categories. In the case of sentiment analysis, the categories would be positive, negative, or neutral. The algorithm behind SVM classifiers aligns almost perfectly with the problem being solved in sentiment analysis, so it seems logical that the SVM classifier would have the best performance.

3) How would you improve the baseline model? (i.e., what text features are most beneficial to the task of sentiment analysis?) Experiment with at least 4 different features that go beyond a language modeling representation.

This baseline model has the potential to be improved in a number of ways. As previously stated, the Bag-of-Words model used here does not consider context. This disregard of context could cause some documents to be incorrectly categorized. For instance, in the bag of words model the word “bad” will likely be associated with negativity. However “bad” could actually indicate a positive connotation in certain contexts. For example, “not bad” typically means that something was satisfactory and in some instances good. In order to be able to identify these nuances, approaches like n-gram models (where n is greater than one) could be attempted in lieu of a bag-of-words. Additionally, the sentiment analysis performed here removed punctuation in

preprocessing, however in some instances punctuation can be an indicator of linguistic features. Consider the sentence, “This ‘wonderful’ car turned out to be a piece of junk” (Girju 2020). In this phrase, the quotation marks indicate that while at some point the car was referred to as wonderful, that is not the current opinion of the writer/speaker. Without considering this punctuation, this phrase might be categorized as positive. For this reason, selectively maintaining certain punctuation also has the potential to improve the accuracy of the baseline sentiment analysis performed here. Another way to improve this baseline model would be to implement dependency parsing. This would help to further inform the classifier about where that sentiment in a text is coming from and at whom it is directed (Girju 2020). Furthermore, this classifier only considers adjectives and adverbs as relevant words, however in some cases verbs and nouns can be indicative of sentiment. It is possible that not filtering out some parts of speech could have a positive effect on this baseline implementation.

4) How does the size of the various feature sets you are experimenting with influence the performance of sentiment analysis? Compare the performance of different feature sets under the same feature selection scenario and machine learning algorithm.

This sentiment analysis was also performed with various sizing for the feature sets. The baseline sizing for feature sets was 5000 words from the lexicon of relevant words. The effect of feature set sizing on performance was then analyzed by using an incremental approach. This was achieved by adding 1000 new word features and measuring the performance on each of the machine learning models. This was repeated to a maximum of 10,000 word features per vector/featureset. The performance of the different classifiers didn’t seem to change much as the

sizing of the feature sets increased. This is not surprising as the vocabulary was ordered based on word frequency. This indicates that having a feature set of the 5,000 most frequently used adjectives and adverbs is appropriate for the scope of this set of data. This also indicates that the features that I chose to add incrementally were not very informative to the model.

Resources

Girju, R. (2020). *Ling 406: Intro to Computational Linguistics*. *LING 406: Intro to Computational Linguistics*. Urbana-Champaign .

Text Classification Using Support Vector Machines (SVM). (2018, October 4). Retrieved from https://monkeylearn.com/text-classification-support-vector-machines-svm/