CS 410: Technology Review
Kristen Fisher [kaf4]
November 5, 2021

# A Review of LingPipe's Sentiment Analysis Technology

LingPipe is a Java based toolkit created for the "linguistic analysis of human language." Alias-i, originally established by Breck Baldwin from the of University of Pennsylvania as Baldwin Language Technologies in 1999, released LingPipe in 2003. LingPipe offers tools for all kinds of text processing including, classification, entity recognition, part-of-speech tagging, clustering, segmentation, and sentiment analysis.

This review focuses on LingPipe's sentiment analysis technology. Sentiment analysis is the classification of an opinioned text object. The opinion of the text is often binary, classified as being "positive" or "negative", but sometimes intermediate categories, such as "neutral" are used. Sentiment analysis is often referred to as opinion mining and can be a very useful tool in consumer research.

LingPipe implements a hierarchical sentiment analysis on movie reviews by breaking up the task into several parts. First, the text must be segmented into sentences. LingPipe has a few methods for this including the SentenceModel interface which segments sentences based on tokens and whitespaces to find sentence boundaries within the text. The next step is to analyze the sentence itself. Sentences which are objective or factual in nature need not be analyzed for sentiment. LingPipe uses a logistic classifier to classify sentence subjectivity. Logistic regression is used to estimate text class (subjective or objective) probabilities based on the features of the input vector, the sentence. The demo for subjectivity analysis on LingPipe's website boasts 92% accuracy in their classifier's evaluation.

Once the sentences of a text have been classified as being objective or subjective, the subjective sentences are then extracted to a string. LingPipe implements a method to reduce the number of sentences to be classified by selecting the top ranked subjective sentences to be utilized in the classifier. As written in the demo, the top 5 sentences with the highest probabilities of being subjective are included automatically, and then up to a total of 20 can be included if the subjectivity score is greater than 50% in favor of a subjective sentence. According to LingPipe's documentation there are multiple ways to configure these thresholds which affect how sentences will be fed into the polarity (sentiment) classifier. The demo for polarity analysis on LingPipe's website boasts 91% accuracy utilizing this hierarchical method of sentiment classification.

A publication from 2004*, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts* by Bo Pang and Lillian Lee at Cornell University, utilized this same hierarchical technique to classify movie reviews comparing the use of support vector machines and Naïve Bayes to train the classifiers. Results were interesting and showed this method of reducing

the text to subjective only context can be quite effective, especially when used with a Naive Bayes sentiment classifier.

In 2013, Vasu Jain from the University of Southern California's Department of Computer Science published the *Prediction of Movie Success using Sentiment Analysis of Tweets*, in which he uses LingPipe to perform sentiment analysis on tweets about movies to predict the movie's success. Accuracy for sentiment analysis in this study was approximately 64%, which is not terrible considering the text objects were tweets and not text documents with proper sentence structure. This article demonstrates potential extensions of LingPipe's sentiment analysis technology. However, when researching the current of state-of-the-art in sentiment analysis, I have found no significant mentions of LingPipe in the past five years. Additional clues that LingPipe sentiment analysis may be a somewhat outdated technology include the fact that the LingPipe blog has not been updated in approximately six years. Internet searches for articles reveal similar results. While LingPipe was likely cutting edge at its release, since this time there have been significant developments in all domains of natural language processing including sentiment analysis.

State-of-the-art sentiment analysis now focuses on new language representation models. One of the top performers is Bidirectional Encoder Representations from Transformers (BERT) developed by Google AI Language. BERT is an autoencoder language model which attempts to reconstruct the original document from a document where words have been "masked." Another is XLNet, which is an autoregressive language model which uses context to predict words. Both methods achieve >95% accuracy on the same IMDb movie review dataset. Additionally, many of LingPipe's competitors have continued to develop and refine their sentiment analysis techniques. Standout competitors in the academic and open-source realm include the General Architecture for Text Engineering (GATE), Natural Language Toolkit for Python, and OpenNLP as well as Lexalytics and Google Cloud Natural Language API in the industrial sector.

## References


http://www.alias-i.com/lingpipe/

http://www.alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html

https://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

http://www.jscse.com/papers/vol3.no3/vol3.no3.46.pdf

http://nlpprogress.com/english/sentiment_analysis.html

https://towardsdatascience.com/what-is-xlnet-and-why-it-outperforms-bert-8d8fce710335