A comparison of machine-learning based sentiment analysis techniques: Support Vector Machines vs. Maximum Entropy

By Maya Subramanian

Sentiment analysis is the “process of analyzing data and classifying it into a category, i.e., positive and negative”[1]. This can be very hard to accomplish with textual data due to the complicated nature of human languages. Several machine-learning techniques currently exist to perform sentiment analysis on textual data and each tries to classify sentiments as accurately as possible. In this review, we will focus on the Support Vector Machine and Maximum Entropy techniques and compare the accuracy of each method on Twitter user data (tweets). We will investigate a comparison of these methods conducted by the Islamic University of Indonesia and published in the International Journal of Advances in Electronics and Computer Science.

First, we will define the two techniques. Support Vector Machines classify text data as either “positive” or “negative” by defining a hyperplane (separator) that differentiates the two classes. Maximum Entropy finds weights for the feature vectors that maximize likelihood of the training data.

The dataset in the study by the Islamic University of Indonesia was obtained from Twitter, as previously mentioned. Researchers gathered tweets about Pertamina, an Indonesian petroleum company, regarding Pertalite, a gasoline fuel sold by them. They used the keyword “Pertalite” in the search bar to get 1411 tweets, which were then reduced to 1099 tweets after omitting totally irrelevant tweets. They aimed to gather information on the public’s response to this product over the months of May, June, and July 2017 using sentiment analysis while analyzing the Support Vector Machine and Maximum Entropy classifier techniques.

The Twitter dataset was preprocessed to extract unnecessary data and normalize the tweets. The researchers tokenized phrases, converted all words to lowercase, filtered out unnecessary data such as usernames, URLs, punctuation, the word “retweet”, and the words “Pertalite”, “Gasoline”, and “Pertamina”, and stemmed the textual data. After cleaning the data, they performed both Support Vector Machine and Maximum Entropy classifiers on this dataset.

The Indonesian researchers found the Support Vector Machine classifier had higher accuracy than the Maximum Entropy classifier by 1.55% in tweets from May-July 2017 overall; however, in both June and July individually, the Maximum Entropy model performed slightly better, with a 2.25% and 2.99% difference in each month, respectively.

They claimed that accuracy levels appeared to be affected by the percentage of data testing positive. In months with a much higher rate of positive tweets, the accuracy of both classifiers was much higher, as well.

Shown below are two tables from Reference 1:

Period	Data Testing	Data Testing Positive	Percentage of Data Testing Positive
May - July	1099	642	58,42
May	62	22	35,48
June	212	139	65,57
July	201	185	92,04

Table 3: Positive data testing percentage

Method	Percentage Correct (%)	Overall Percentage (%)
Mei - Juli	*Maxent*	66,42
	SVM	67,97
Mei	*Maxent*	62,90
	SVM	67,74
Juni	*Maxent*	86,32
	SVM	82,07
Juli	*Maxent*	98,01
	SVM	95,02

Tabel 4: Overall percentage

Based on their findings, we can see that both the Support Vector Machine and Maximum Entropy classifiers have similar accuracy levels and are highly competitive with each other, with SVM performing slightly better overall. It is interesting that according to the researchers, the increase in the overall percentage of positive tweets resulted in a higher accuracy in both these classifiers.

I wonder if the researchers were actually *incorrect* with this assumption, in a classic case of correlation vs. causation. I believe the reason for the classifiers' accuracy increase was due to the increase in data as a whole. I did not find empirical evidence to this claim in their findings as to why it is not just simply the increase of data per month that has increased the accuracy rather than the increase of positive tweets only. As we can observe in the table above, a mere 62 tweets were used in May 2017. 212 tweets were used in June 2017, and 201 tweets were used in July 2017. The accuracy shot up from ~65% or so for both classifiers in May 2017 to ~84% in June 2017 and ~96% in July 2017. In the machine learning world, it is common knowledge that the more data, the better the classifier.

The Indonesian researchers also did not mention if they fine tuned parameters for each month, or if the parameters remained the same for all three months of data, but this can play an impact on the overall accuracy as well.

To further investigate the fact that more data = higher accuracy, I researched a different study by a professor and student in the Department of Computer Science and Engineering at the Ambedkar Institute of Advanced Communication Technologies & Research. They conducted sentiment analysis on Twitter data as well using the Support Vector Machine method. These researchers gathered 1100 tweets in one dataset with 1100 and 15,662 tweets in another dataset. Similarly to the Indonesian researchers, these researchers pre-processed their data. They decided to use the unigram model to analyze one word at a time in the SVM classifier.

TABLE V
ACCURACY COMPARISON OF THE CLASSIFIER

Dataset	Accuracy of SVM
Twitter dataset 1 of 1000 tweets	50.85%
Twitter dataset 2 of 10,662 tweets	71.60%

These researchers found the accuracy of the first dataset to be 50.85% and the second dataset to be 71.60%. As previously mentioned, classifiers perform better with a larger dataset – this is also shown in this case with a ~20% jump in accuracy in the dataset with a much higher volume of tweets. They even stated in their conclusion, "We found that size of a dataset greatly impacts on the accuracy of the classifier."

Both the Maximum Entropy and SVM models are shown to comparably predict the sentiment of tweets with similar accuracy levels. There are several factors that can improve these classifiers, including the size of the dataset, as I've shown in this tech review.

References:

1) **Comparison of Maximum Entropy and Support Vector Machine Methods For Sentiment Analysis of Pertalite Product Through Twitter Social Network.** http://www.iraj.in/journal/journal_file/journal_pdf/12-430-152034023210-14.pdf

2) **Sentiment Analysis on Twitter Data using Support Vector Machine.** http://www.ijsrcsams.com/images/stories/Past_Issue_Docs/ijsrcsamsv7i3p38.pdf