

# LING 575: Intermediate Project Report

**Claire Jaja**

University of Washington  
Seattle, WA  
cjaja@uw.edu

**Andrea Kahn**

University of Washington  
Seattle, WA  
amkahn@uw.edu

## Abstract

## 1 Introduction

Sentiment analysis techniques are typically developed on English text. However, there is a proliferation of text in other languages as well, which could benefit from sentiment analysis. Current approaches to sentiment analysis in languages other than English often involve automatically translating the text into English, which is likely to introduce errors, as well as increasing runtimes. Some techniques rely on manually developed resources which are only available in English; however, others rest on machine learning algorithms that can easily be applied to other languages. In this paper, we propose an approach to evaluate the effectiveness of these techniques when transferred directly to other languages.

## 2 Related Work

## 3 Data

For our experiments, we are using comparable corpora in multiple languages.

Firstly, we use IMDb data from the Pang/Lee ACL 2004 polarity dataset v2.0, which consists of 1000 positive and 1000 negative pre-processed reviews. For a comparable dataset in another language, we use CorpusCine Reviews (Cruz Mata, 2011) - a collection of 3,878 movie reviews written in Spanish from the muchocine.net web page. Each review has a rating between one (most negative) and five (most positive) stars. For our purposes, we discard the three star reviews and classify the one and two star reviews as negative and the four and five star reviews as positive. This leaves us with 1,274 negative reviews and 1,351 positive reviews.

Additionally, we have a corpus of 1,590 English quotations from newspaper articles annotated for

polarity and a comparable corpus of 2,387 German quotations from newspaper articles; both of these were annotated based on the same annotation criteria.

## 4 Approach

We employ the Mallet toolkit to train and test polarity classifiers on these two datasets. We plan to test a variety of classification algorithms and features, in order to determine if certain ones provide better results for one language or another. Additionally, we plan to do a linguistic error analysis, to tease apart the impact of the language differences (in this vein, Andrea plans to use this course as a linguistics elective, while Claire is using it as a computational linguistics elective).

## 5 Results

We have implemented a baseline system on both movie review datasets using a MaxEnt classifier and unigram bag of word features. For each dataset, we use 90% of the reviews for training and 10% for testing; this gives us a training set of 900 positive and 900 negative reviews for the IMDb corpus and 1,215 positive and 1,146 negative reviews for the CorpusCine corpus. There are 100 positive and 100 negative reviews in the IMDb test set and 136 positive and 128 negative reviews in the CorpusCine test set. The results for this system are presented below.

Classifier	features	IMDb	CorpusCine
MaxEnt	unigram	88.00%	83.71%

Table 1: Baseline results. Test accuracy for IMDb and CorpusCine.

label	negative	positive
negative	86	14
positive	10	90

Table 2: IMDb confusion matrix. Row = true,

*column = predicted.*

<b>label</b>	<i>negative</i>	<i>positive</i>
<i>negative</i>	103	25
<i>positive</i>	18	118

*Table 3: CorpusCine confusion matrix. Row = true, column = predicted.*

## 6 Discussion

Our next steps include cutting back the CorpusCine train and test sets to have the same number of reviews as the IMDb train and test sets to make the results as directly comparable as possible. To this end, we will also compare review lengths (i.e. average, minimum, and maximum number of tokens and types). After that, we hope to implement n-fold cross validation of our results (most likely with n=10). Once our datasets are fully established, we will test different classification algorithms and features, as well as external resources (e.g. polarity lexicons, both native and translated).

## 7 Conclusion