Multilingual Sentiment Analysis of IMDb Movie Reviews

**Introduction**

Sentiment Analysis is a subfield of Natural Language Processing (NLP) that involves the training of learners that can quantize and extract sentiment values (i.e. positive, negative, or neutral opinion) from text data. Sentiment analysis is one of many aspects of the broader domain of text mining. Sentiment analysis can have varying granularity: it can be performed at the document level, sentence level, or even sub-sentence level [1].

Multilingual Sentiment Analysis extends the concepts of sentiment analysis by introducing multiple natural languages. Multilingual sentiment analysis is highly valuable for product analytics, market research, social media monitoring, and customer service. For example, a book publishing company wants to determine whether they should reprint an out-of-print book and the international markets where they should prioritize shipping. After scraping reviews from online review sites and forums, they find that the book is well-received and in-demand in Asian markets; however, European readers elicited a lukewarm response. Armed with this information, the company diverts advertising and shipping resources to Asia.

**Preprocessing of IMDb Movie Reviews**

The Internet Movie Database (IMDb), contains reviews for millions of movie and television titles. Using lexicon-based sentiment analysis and a simple machine translation approach, we can test the variances in sentiment that may occur amongst various natural languages. Out of 25,000 available movie reviews, a small partition of 60 reviews with pre-defined sentiments (30 positive, 30 negative) were chosen semi-randomly. From this partition, three sets were created with the aid of machine translation: the original English movie reviews, English-to-Spanish reviews, and English-to-Spanish-to-English reviews.

%Read English samples (30 positive, 30 negative IMDb reviews)

file\_samplesENG = fopen('SamplesENG.txt');

C = textscan(file\_samplesENG,'%s','Delimiter','\n');

samplesENG = string(C{1});

%Read Machine-translated Spanish samples

file\_samplesESP = fopen('SamplesENGtoESP.txt');

C = textscan(file\_samplesESP,'%s','Delimiter','\n');

samplesESP = string(C{1});

%Read ENG->ESP->ENG samples

file\_samples\_final = fopen('SamplesENGtoESPtoENG.txt');

C = textscan(file\_samples\_final,'%s','Delimiter','\n');

samples\_final = string(C{1});

One of the most vital components of any manner of text mining is preprocessing of the input text data. In sentiment analysis, “Dirty” text data can have adverse effects on the output sentiment and classifier values. Some issues of input text data could have a lot of noise, spam/adversarial responses, and/or blank responses. This issue was encountered early on when working with the IMDb dataset. Several entries included the <br> HTML tag, likely an unresolved error from the web scraper used to collect the data. These tags were removed from the affected entries before any other form of preprocessing was applied.

Word clouds use the frequency of tokens within a document to display the most relevant information. It’s important that we have visual evidence of the changes applied to the datasets before and after preprocessing.

%Wordcloud visualization

%Wordclouds without preprocessing

figure(1);

subplot(1,2,1);

wordcloud(samplesENG);

title("Wordcloud for English Samples before preprocessing");

figure(2);

subplot(1,2,1);

wordcloud(samplesESP);

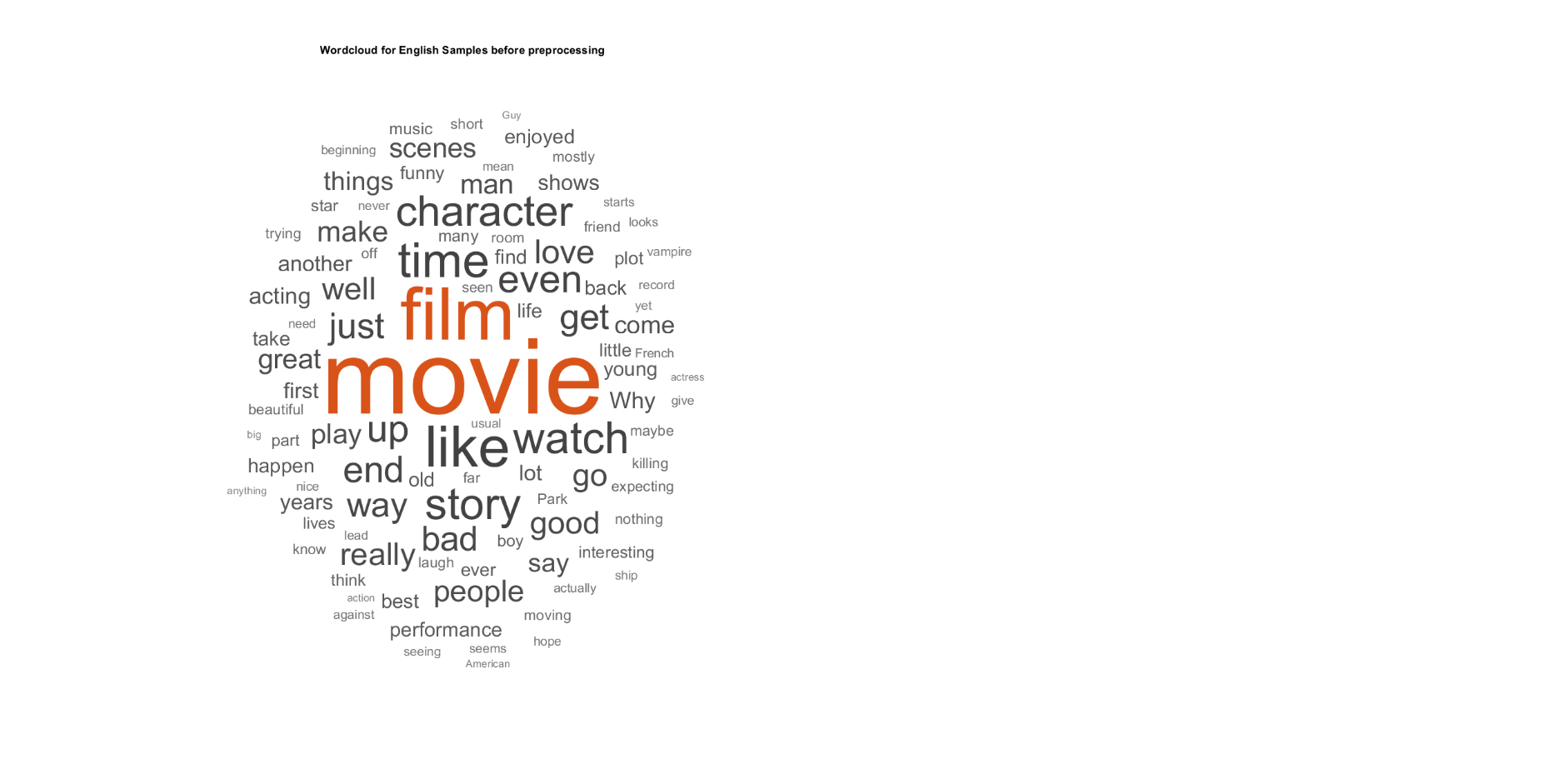
title("Wordcloud for ENG->ESP Samples before preprocessing");

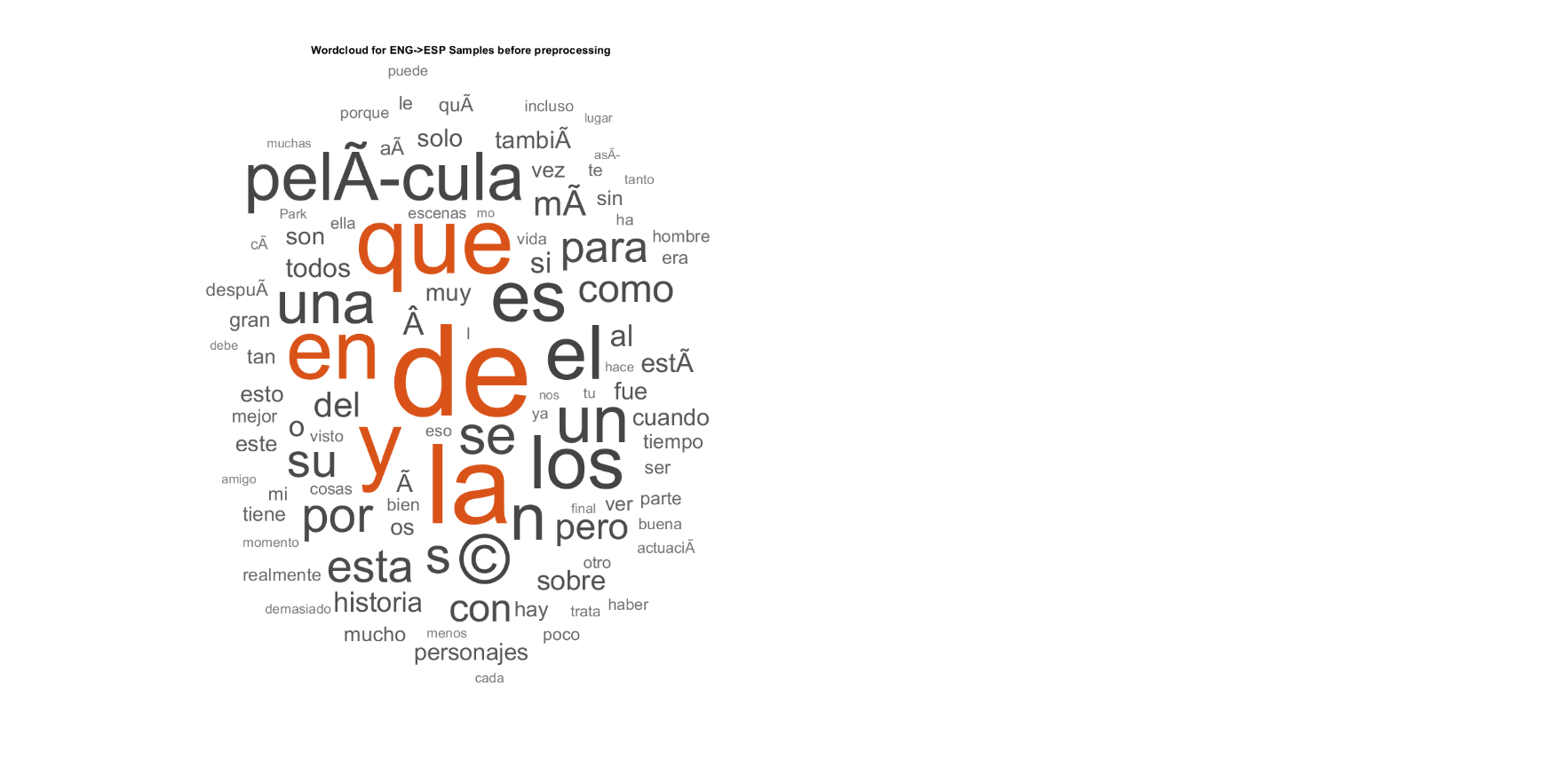
figure(3);

subplot(1,2,1);

wordcloud(samples\_final);

title("Wordcloud for ENG->ESP->ENG Samples before preprocessing");







One of the things we notice immediately is that the ENG->ESP word cloud is dominated by conjunctions. Fortunately, the ENG->ESP->ENG frequencies appear to be very similar to the original English frequencies.

Preprocessing the data and generating clean documents for sentiment analysis is straightforward. We start by converting all text data to lowercase. Next, the lines of each document are split into tokens.

130 tokens: this movie is over hyped ! ! i am sad to say that i manage to watch the first 15 minutes of this movie and anything beyond that , i will have to force myself real hard to sit down and watch the rest of the movie . it's totally stupid and very fake . the robot in the movie looks like a man wearing those steel suit and the acting is really bad especially the one playing the character alien . he is totally annoying ! ! don't waste your money watching this sequel to the popular gen-x cops . i'd rather sleep or spend my money on some other things rather than watching this movie . 1 out of 10 . if possible , i'd give 0 .

Now, we can remove punctuation and *stop words*. Stop words are short words such as “the”, “a”, “an”, “in”, etc. that are useless for determining sentiment. We can also eliminate words that are too short or too long, based on user-defined criteria. We chose to remove words shorter than two characters or longer than twenty-five characters. The penultimate step is to stem and lemmatize the input data. This process may also be referred to as *suffix stripping*. For example, “playing” is transformed into its root, “play”. Finally, we can remove any blank responses that may exist and words that appear infrequently, e.g. less than twice per review.  
  
57 tokens: movie hyped sad say manage watch first minutes movie anything beyond force myself real hard sit down watch rest movie totally stupid fake robot movie looks like man wearing steel suit acting really bad especially playing character alien totally annoying waste money watching sequel popular genx cops rather sleep spend money things rather watching movie possible give

After preprocessing, our word cloud visualizations show heightened frequencies for the descriptive values we are seeking. In the future, a more sophisticated algorithm should be used to target and remove specific words in the dataset that appear frequently but have little sentiment value (e.g. “film” and “movie”).

%--------------------------------------

%Preprocessing

%Change all text data to lowercase

cleanTextData1 = lower(samplesENG);

cleanTextData2 = lower(samplesESP);

cleanTextData3 = lower(samples\_final);

cleanTextData1(1:10)

%Tokenize data

cleanDocuments1 = tokenizedDocument(cleanTextData1);

cleanDocuments2 = tokenizedDocument(cleanTextData2);

cleanDocuments3 = tokenizedDocument(cleanTextData3);

cleanDocuments1(1:10)

%Remove punctuation

cleanDocuments1 = erasePunctuation(cleanDocuments1);

cleanDocuments2 = erasePunctuation(cleanDocuments2);

cleanDocuments3 = erasePunctuation(cleanDocuments3);

%Remove stop words

cleanDocuments1 = removeStopWords(cleanDocuments1);

cleanDocuments2 = removeStopWords(cleanDocuments2);

cleanDocuments3 = removeStopWords(cleanDocuments3);

%Remove words that are too short or too long

cleanDocuments1 = removeShortWords(cleanDocuments1,2);

cleanDocuments1 = removeLongWords(cleanDocuments1,25);

cleanDocuments2 = removeShortWords(cleanDocuments2,2);

cleanDocuments2 = removeLongWords(cleanDocuments2,25);

cleanDocuments3 = removeShortWords(cleanDocuments3,2);

cleanDocuments3 = removeLongWords(cleanDocuments3,25);

cleanDocuments1(1:10)

%Lemmatize the words

cleanDocuments1 = addPartOfSpeechDetails(cleanDocuments1);

cleanDocuments1 = normalizeWords(cleanDocuments1,'Style','lemma');

cleanDocuments2 = addPartOfSpeechDetails(cleanDocuments2);

cleanDocuments2 = normalizeWords(cleanDocuments2,'Style','lemma');

cleanDocuments3 = addPartOfSpeechDetails(cleanDocuments3);

cleanDocuments3 = normalizeWords(cleanDocuments3,'Style','lemma');

cleanDocuments1(1:10);

%Remove infrequent words

bag1 = bagOfWords(cleanDocuments1);

cleanBag1 = removeInfrequentWords(bag1,2);

bag2 = bagOfWords(cleanDocuments2);

cleanBag2 = removeInfrequentWords(bag2,2);

bag3 = bagOfWords(cleanDocuments3);

cleanBag3 = removeInfrequentWords(bag3,2);

%Remove empty documents

cleanBag1 = removeEmptyDocuments(cleanBag1);

cleanBag2 = removeEmptyDocuments(cleanBag2);

cleanBag3 = removeEmptyDocuments(cleanBag3);

figure(4);

subplot(1,2,2);

wordcloud(cleanBag1);

title("Wordcloud for English Samples after preprocessing");

figure(5);

subplot(1,2,2);

wordcloud(cleanBag2);

title("Wordcloud for ENG->ESP Samples after preprocessing");

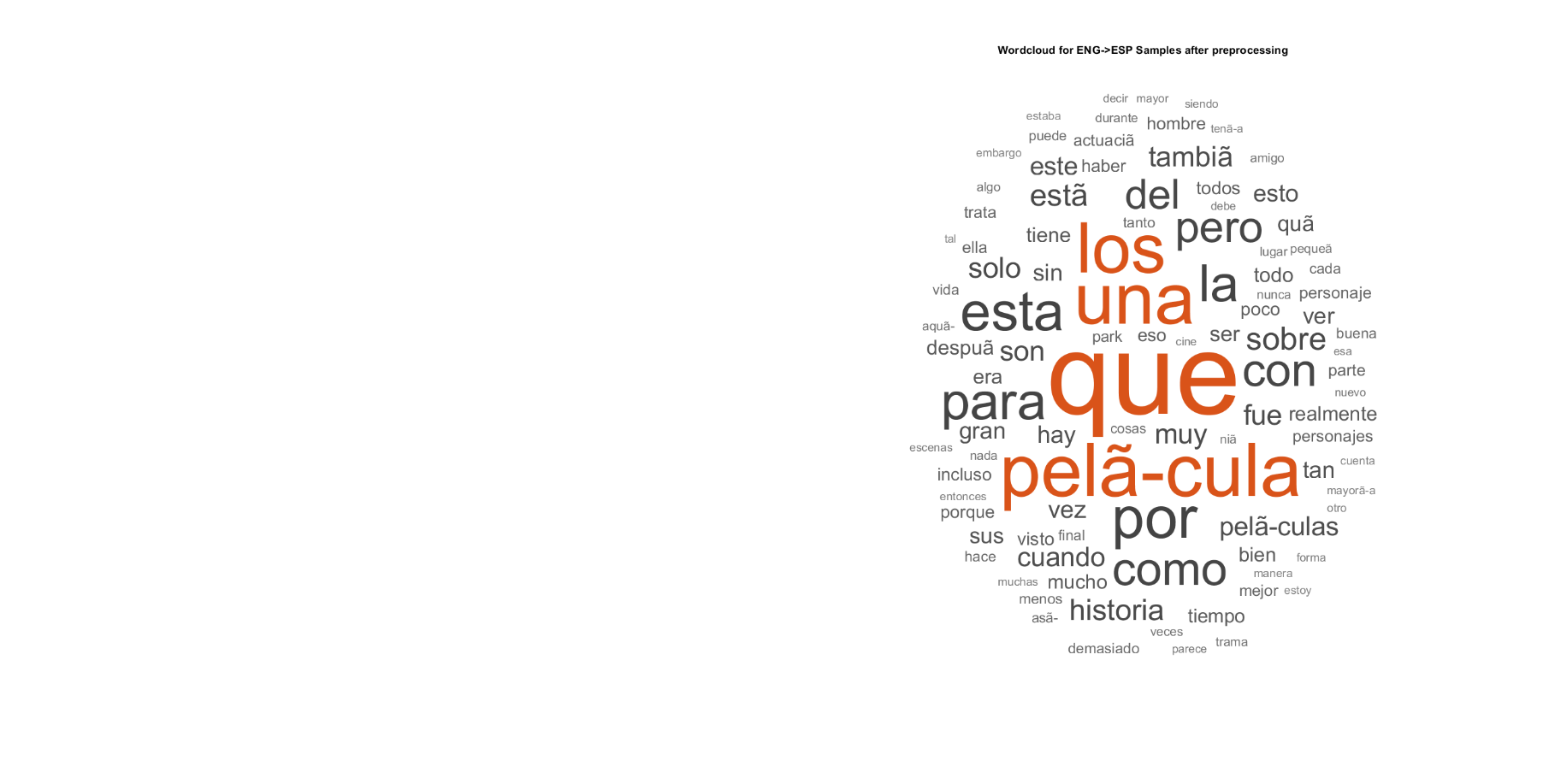
figure(6);

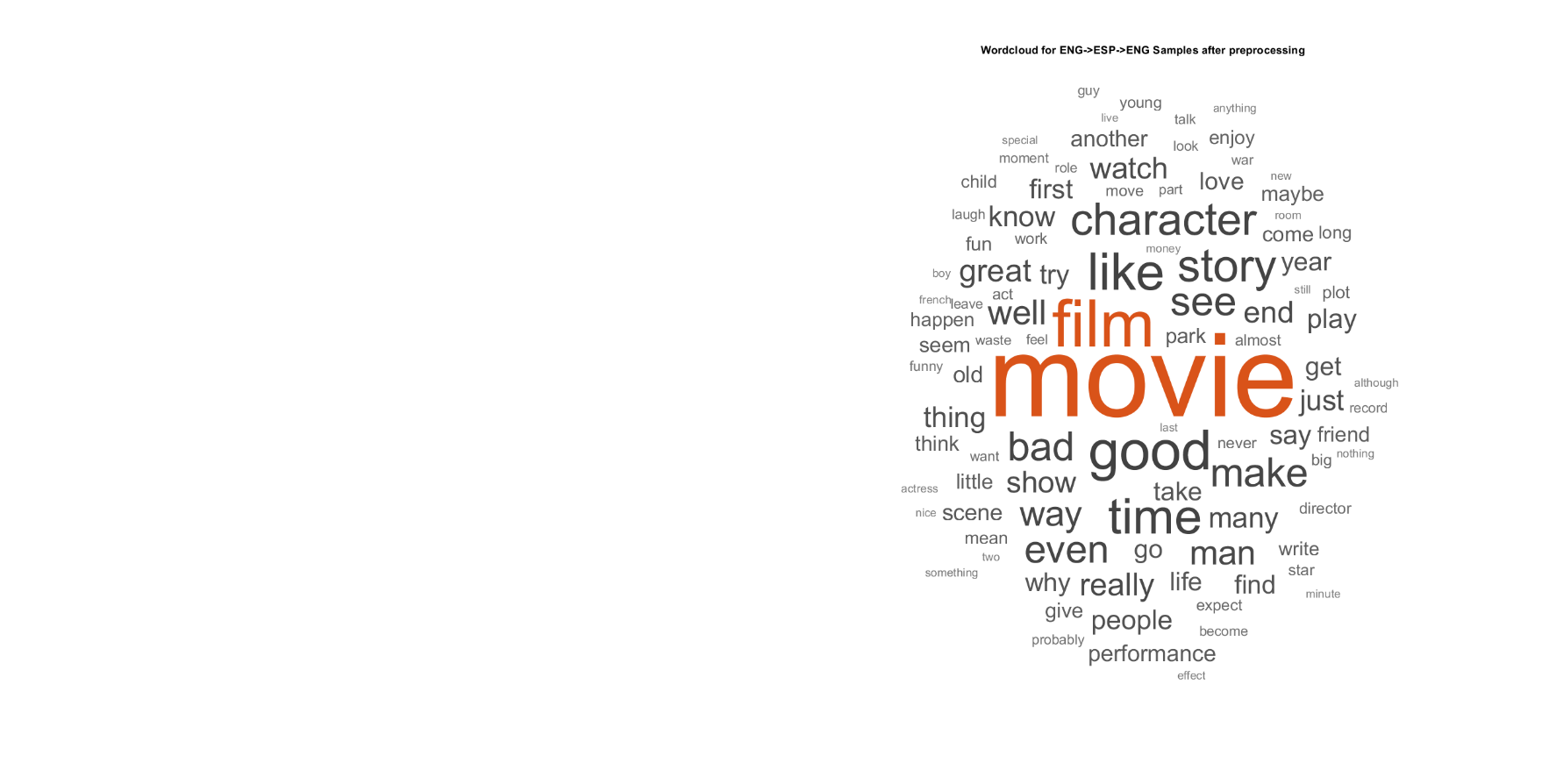
subplot(1,2,2);

wordcloud(cleanBag3);

title("Wordcloud for ENG->ESP->ENG Samples after preprocessing");







**Application of Sentiment Classifier**

Our lexicon-based sentiment classifier used the free and pre-trained fastText word embedding [2], as well as a publicly available English opinion lexicon [3]. Our classifier was trained using Word2Vec (a shallow neural net designed to process text), and a Support Vector Machine (SVM).

%load pretrained word embedding

emb = fastTextWordEmbedding;

%Load Opinion Lexicon

data = readLexicon;

%Train the Sentiment Classifier

idx = ~isVocabularyWord(emb,data.Word);

data(idx,:) = [];

numWords = size(data,1);

cvp = cvpartition(numWords,'HoldOut',0.1);

dataTrain = data(training(cvp),:);

dataTest = data(test(cvp),:);

wordsTrain = dataTrain.Word;

XTrain = word2vec(emb,wordsTrain);

YTrain = dataTrain.Label;

%Fit model to an SVM

mdl2 = fitcsvm(XTrain,YTrain);

wordsTest = dataTest.Word;

XTest = word2vec(emb,wordsTest);

YTest = dataTest.Label;

[YPred,scores] = predict(mdl2,XTest);

%Plot confusion matrix

figure

confusionchart(YTest,YPred);

%Wordcloud visualization of trained pos./neg. sentiment

figure

subplot(1,2,1)

idx = YPred == "Positive";

wordcloud(wordsTest(idx),scores(idx,1));

title("Predicted Positive Sentiment")

subplot(1,2,2)

wordcloud(wordsTest(~idx),scores(~idx,2));

title("Predicted Negative Sentiment")

%This function only needs to run once in the MATLAB workspace -- for importing the opinion lexicon.

function data = readLexicon

% Read positive words

fidPositive = fopen(fullfile('opinion-lexicon-English','positive-words.txt'));

C = textscan(fidPositive,'%s','CommentStyle',';');

wordsPositive = string(C{1});

% Read negative words

fidNegative = fopen(fullfile('opinion-lexicon-English','negative-words.txt'));

C = textscan(fidNegative,'%s','CommentStyle',';');

wordsNegative = string(C{1});

fclose all;

% Create table of labeled words

words = [wordsPositive;wordsNegative];

labels2 = categorical(nan(numel(words),1));

labels2(1:numel(wordsPositive)) = "Positive";

labels2(numel(wordsPositive)+1:end) = "Negative";

data = table(words,labels2,'VariableNames',{'Word','Label'});

end

function [documents] = preprocessReviews(textData)

% Convert the text data to lowercase.

cleanTextData = lower(textData);

% Tokenize the text.

documents = tokenizedDocument(cleanTextData);

% Erase punctuation.

documents = erasePunctuation(documents);

% Remove a list of stop words.

documents = removeStopWords(documents);

%Remove long and short words

documents = removeShortWords(documents,2);

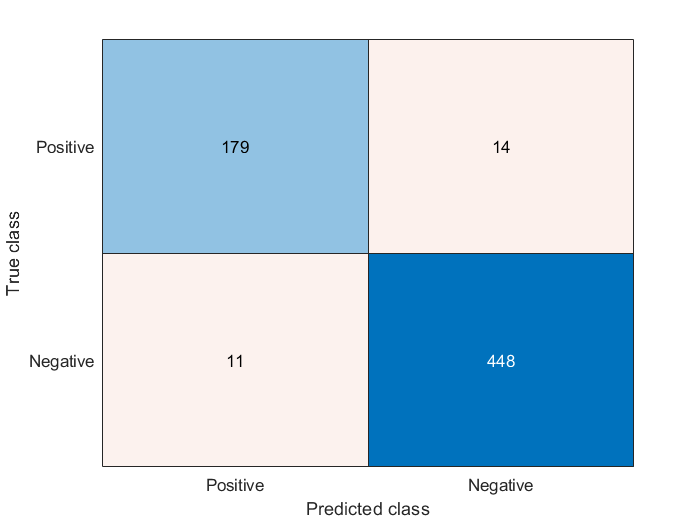
documents = removeLongWords(documents,25);

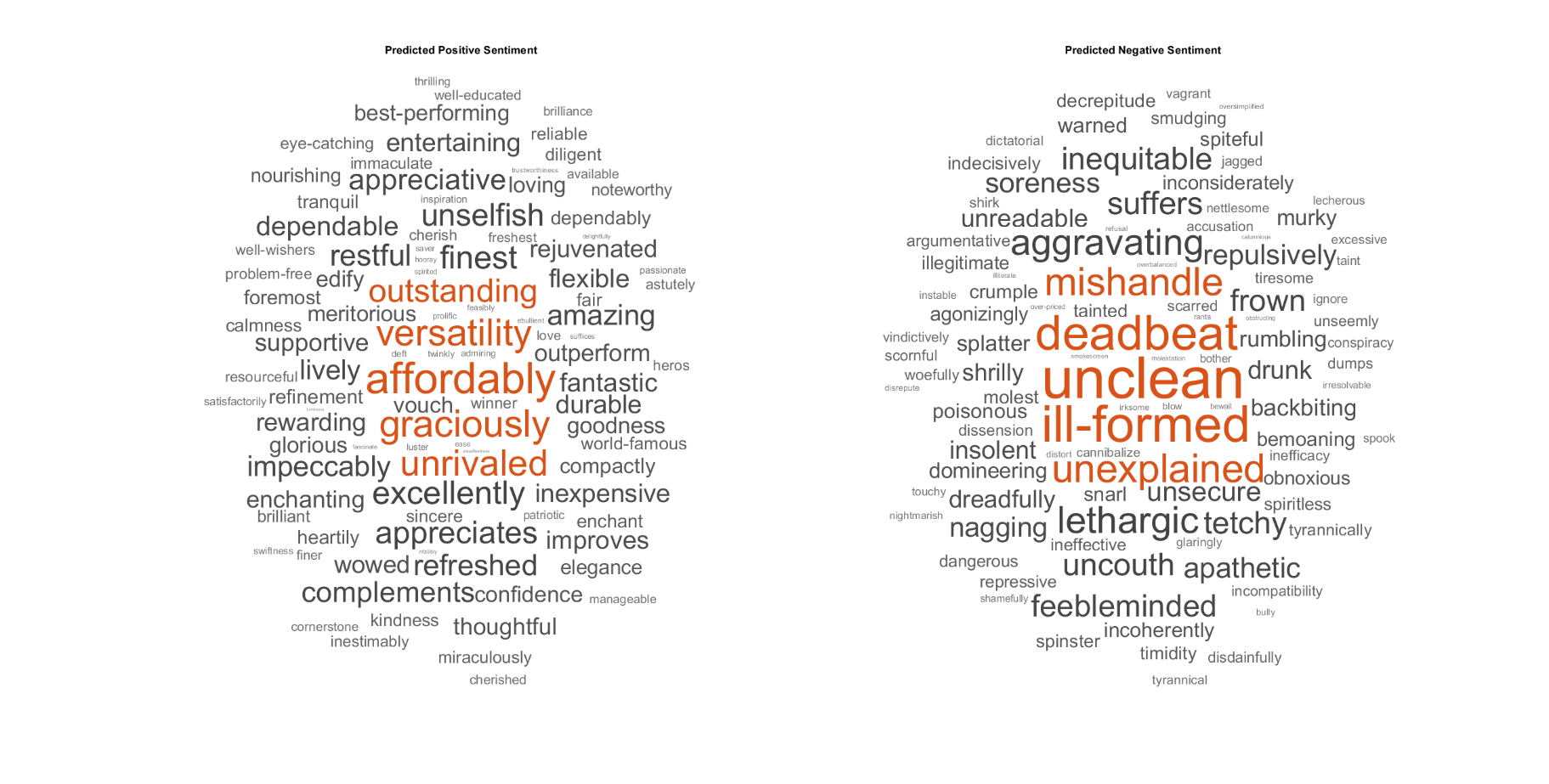
%Stem and lemmatize the documents

documents = addPartOfSpeechDetails(documents);

documents = normalizeWords(documents,'Style','lemma');

end





**Results of Sentiment Analysis**

Now that the sentiment classifier is trained, we can calculate the sentiment values of our datasets.

%Calculate Sentiment for a given sample and/or set of samples

textData1 = samplesENG;

textData2 = samplesESP;

textData3 = samples\_final;

documents1 = preprocessReviews(textData1);

idx = ~isVocabularyWord(emb,documents1.Vocabulary);

documents1 = removeWords(documents1,idx);

idx = 1:60;

for i = 1:numel(idx)

words = string(documents1(idx(i)));

vec = word2vec(emb,words);

[~,scores] = predict(mdl2,vec);

sentimentScore1(i) = mean(scores(:,1));

end

[sentimentScore1' textData1(idx)]

documents2 = preprocessReviews(textData2);

idx = ~isVocabularyWord(emb,documents2.Vocabulary);

documents2 = removeWords(documents2,idx);

idx = 1:60;

for i = 1:numel(idx)

words = string(documents2(idx(i)));

vec = word2vec(emb,words);

[~,scores] = predict(mdl2,vec);

sentimentScore2(i) = mean(scores(:,1));

end

[sentimentScore2' textData2(idx)]

documents3 = preprocessReviews(textData3);

idx = ~isVocabularyWord(emb,documents3.Vocabulary);

documents3 = removeWords(documents3,idx);

idx = 1:60;

for i = 1:numel(idx)

words = string(documents3(idx(i)));

vec = word2vec(emb,words);

[~,scores] = predict(mdl2,vec);

sentimentScore3(i) = mean(scores(:,1));

end

[sentimentScore3' textData3(idx)]

ans =

60×2 string array

"-0.26386" "For years, I've been a big fan of Park's w…"

"-0.011227" "Might end up being the biggest disappointm…"

"-0.42408" "I only came here to check Terror Hospital …"

"-0.19299" "I and a friend rented this movie. We both …"

"-0.03216" "A sprawling, overambitious, plotless comed…"

"-0.2048" "Ben, (Rupert Grint), is a deeply unhappy a…"

"-0.57361" "This movie is over hyped!! I am sad to say…"

"0.044543" "My god...i have not seen such an awful mov…"

"-0.18955" "and I still don't know where the hell this…"

"-0.31369" "I have to admit that I am disappointed aft…"

"0.30407" "There must be an error. This movie belongs…"

"-0.026869" "I don't know any idiotic rock'n'roll clich…"

"0.14817" "They missed up the film when the tried to …"

"-0.10972" "My wife and I saw this in the theater when…"

"-0.47277" "The preposterous premise of this flick has…"

"-0.0098485" "While credited as a Tom and Jerry cartoon,…"

"-0.87903" "Lame B-horror that takes itself too damn s…"

"-0.37972" "How better to describe it than scuzzy crim…"

"-0.48265" "I think i would rather have my piles clipp…"

"0.16184" "I was expecting a documentary covering the…"

"-0.021513" "I couldn't believe how bad this film was, …"

"-0.66953" "Okay, so when a friend of mine told me he …"

"-0.94706" "John Leguizamo must have been insane if he…"

"-0.81427" "Relesed from Troma (which is my favorite m…"

"0.18031" "If you first saw this movie with Mary of t…"

"0.35058" "After a string of successful 'a man and hi…"

"0.31864" "Why is Guy working for Buddy? Probably bec…"

"0.1336" "If I'm to like a movie, I need to care abo…"

"0.6779" "Ravi Chopra wrote this film 40 years back,…"

"-0.56076" "Another movie from Swedish hillbilly count…"

"0.28694" "As others that have commented around the w…"

"0.26637" "Anything Park Chan-wook creates is guarant…"

"0.82568" "Another great musical from Hollywoods Gold…"

"0.12859" "i have lost count as to how many times i h…"

"0.78557" "Although it doesn't seem very promising fo…"

"-0.048573" "I saw this movie last night on HBO & didn'…"

"-0.49602" "Meant to be some sort of a social commenta…"

"0.82942" "I watched this film on ITV and I enjoyed i…"

"0.70832" "President Harry S. Truman once said that t…"

"-0.01515" "When I first heard about "Greek," I figure…"

"0.97085" "A delightful story about two evacuees, has…"

"-0.097972" "If all movies had to be destroyed and only…"

"0.34208" "I happened upon this by chance. I was at m…"

"0.056655" "This movie was made by Daiei Studios, know…"

"0.20485" "This is a beautiful, yet simple movie abou…"

"-0.59114" "If you loved the 1993 (erotic, sci-fiction…"

"0.2992" "I'm a large scarred heterosexual male ex-b…"

"0.37335" "This short subject gathered kudos from all…"

"0.45431" "this movies is really special ! it's about…"

"0.63226" "I would give this movie high marks for the…"

"0.34548" "I recalled watching this program as a youn…"

"-0.11044" "First of all, ignore the comment about how…"

"-0.064522" "I don't see what the problem is with SOME …"

"0.33927" "William Faulkner was one of the American w…"

"0.89423" "As many of today's movies are guilty of, t…"

"0.039523" "A lot of people say the end did not make s…"

"0.13638" "This movie has one of the best club scenes…"

"0.45537" "This is not your typical Indian film. Ther…"

"-0.12337" "SPOILERS AHEAD For the first ten minutes o…"

"0.10212" "Watching "Plots with a View" (called "Unde…"

ans =

60×2 string array

"-0.048132" "Durante aÃ±os, he sido un gran fanÃ¡tico d…"

"0.47493" "PodrÃ­a terminar siendo la mayor decepciÃ³…"

"-0.013138" "Solo vine aquÃ­ para buscar en el Hospital…"

"-0.19996" "Yo y un amigo alquilamos esta pelÃ­cula. A…"

"-0.53194" "Una comedia en expansiÃ³n, demasiado ambic…"

"0.35207" "Ben, (Rupert Grint), es un adolescente pro…"

"-0.08314" "Esta pelÃ­cula estÃ¡ sobre promocionada! M…"

"0.0024012" "Dios mÃ­o ... no he visto una pelÃ­cula ta…"

"-0.60343" "No malgastes tu dinero viendo esta secuela…"

"0.13742" "Â¿Y todavÃ­a no sÃ© a dÃ³nde diablos va es…"

"0.25082" "Tengo que admitir que estoy decepcionado d…"

"0.43039" "Debe haber un error. Esta pelÃ­cula perten…"

"0.014162" "No conozco ningÃºn clichÃ© de rock and rol…"

"-0.32039" "Se perdieron la pelÃ­cula cuando intentaro…"

"0.20072" "Mi esposa y yo vimos esto en el teatro cua…"

"-0.034236" "La premisa absurda de esta pelÃ­cula tiene…"

"0.19222" "Si bien se le atribuye una caricatura de T…"

"0.13358" "Lame B-horror, que se toma demasiado en se…"

"0.40395" "Â¿QuÃ© mejor manera de describirlo que los…"

"0.017757" "Creo que preferirÃ­a que mis pilas se reco…"

"0.32648" "Esperaba un documental que cubriera la era…"

"0.39527" "No podÃ­a creer lo mala que era esta pelÃ­…"

"-0.060373" "De acuerdo, cuando un amigo mÃ­o me dijo q…"

"-0.51378" "John Leguizamo debe haber estado loco si p…"

"0.2709" "Relesed de Troma (que es mi compaÃ±Ã­a de …"

"0.053331" "Si viste esta pelÃ­cula por primera vez co…"

"0.39137" "DespuÃ©s de una serie de exitosas pelÃ­cul…"

"0.10661" "Â¿Por quÃ© estÃ¡ trabajando Guy para Buddy…"

"0.29749" "Si me gusta una pelÃ­cula, debo preocuparm…"

"0.51771" "Ravi Chopra escribiÃ³ esta pelÃ­cula hace …"

"0.098604" "Otra pelÃ­cula del paÃ­s de las colinas su…"

"0.19239" "Como otros que han comentado en la web ...…"

"0.086075" "Todo lo que crea Park Chan-wook estÃ¡ gara…"

"0.39485" "Â¡Otro gran musical de la Edad de Oro de H…"

"0.41926" "He perdido la cuenta de cuÃ¡ntas veces he …"

"0.11104" "Aunque no parece muy prometedor para un tr…"

"0.47012" "Vi esta pelÃ­cula anoche en HBO y no esper…"

"-0.12349" "Significada como una especie de comentario…"

"0.70401" "Vi esta pelÃ­cula en ITV y la disfrutÃ© mu…"

"0.30942" "El presidente Harry S. Truman dijo una vez…"

"0.22107" "Cuando escuchÃ© por primera vez sobre "Gri…"

"0.2576" "Una historia encantadora sobre dos evacuad…"

"0.15458" "Si todas las pelÃ­culas debÃ­an ser destru…"

"0.16798" "Me encontrÃ© con esto por casualidad. Esta…"

"-0.055614" "Esta pelÃ­cula fue hecha por Daiei Studios…"

"0.21953" "Esta es una pelÃ­cula hermosa, pero simple…"

"0.054103" "Si te encantÃ³ la pelÃ­cula de 1993 (erÃ³t…"

"0.18233" "Soy un gran ex-portero masculino, ex-jugad…"

"0.072978" "Este breve tema reuniÃ³ elogios de todo ti…"

"0.66958" "Â¡Esta pelÃ­cula es realmente especial! se…"

"0.56087" "Yo le darÃ­a a esta pelÃ­cula altas califi…"

"0.073104" "RecordÃ© haber visto este programa cuando …"

"0.10431" "En primer lugar, ignora el comentario sobr…"

"-0.018864" "No veo cuÃ¡l es el problema con ALGUNAS pe…"

"-0.046982" "William Faulkner fue uno de los escritores…"

"0.40065" "Como muchas de las pelÃ­culas de hoy son c…"

"-0.15909" "Â¡Esta es sin duda la pelÃ­cula mÃ¡s emoci…"

"0.25907" "Esta no es tu tÃ­pica pelÃ­cula india. Hay…"

"0.041805" "SPOILERS ADELANTE Durante los primeros die…"

"0.30675" "Al ver "Plots with a View" (llamado "Under…"

ans =

60×2 string array

"-0.14912" "For years, I've been a big fan of Park's w…"

"0.044614" "It could end up being the biggest disappoi…"

"-0.35951" "I only came here to look for an alternativ…"

"-0.22731" "Me and a friend rent this movie. We both d…"

"0.36909" "A comedy in expansion, too ambitious and w…"

"-0.039528" "Ben, (Rupert Grint), is a deeply unhappy t…"

"-0.44192" "This movie is about promoted! It saddens m…"

"0.15985" "My God ... I have not seen such a horrible…"

"-0.12431" "Do not waste your money watching this sequ…"

"-0.076335" "And I still do not know where the hell thi…"

"-0.24498" "I have to admit that I am disappointed aft…"

"0.70434" "It must be an error. This movie belongs to…"

"-0.055692" "I do not know any cliche of rock and roll …"

"-0.22582" "They missed the movie when they tried to u…"

"-0.12962" "My wife and I saw this in the theater when…"

"-0.28677" "The absurd premise of this film has to do …"

"0.016301" "Although a cartoon of Tom and Jerry is att…"

"-0.4554" "Lame B-Horror, who takes the subject too s…"

"-0.23308" "What better way to describe it than the fu…"

"-0.52338" "I think I'd prefer that my piles be cut ou…"

"0.097302" "I was expecting a documentary covering the…"

"0.092676" "I could not believe how bad this movie was…"

"-0.50787" "Okay, when a friend of mine told me that h…"

**"-0.90649**" "John Leguizamo must have been crazy if he …"

"-0.55078" "Relesed de Troma (which is my favorite mov…"

"0.17787" "If you saw this movie for the first time w…"

"0.35536" "After a series of successful films of "a m…"

"0.37467" "Why is Guy working for Buddy? Probably bec…"

"-0.27545" "If I like a movie, I have to worry about t…"

"0.69334" "Ravi Chopra wrote this film 40 years ago, …"

"-0.50768" "Another film of the country of the Swedish…"

"0.24717" "Like others who have commented on the web …"

"0.20108" "Everything that Park Chan-wook creates is …"

"**0.91296**" "Another great musical of the Golden Age of…"

"0.11182" "I've lost count of how many times I've see…"

"**0.79108**" "Although it does not seem very promising f…"

"-0.027685" "I watched this movie last night on HBO and…"

"-0.33677" "Meant as a kind of social commentary on th…"

"0.81553" "I saw this movie at ITV and I enjoyed it a…"

"0.68485" "President Harry S. Truman once said that t…"

"0.25941" "When I first heard about "Greek", I though…"

"**1.1135**" "A lovely story about two evacuees, it has …"

"-0.038391" "If all the films had to be destroyed and o…"

"0.48892" "I came across this by chance. I was at my …"

"0.081279" "This movie was made by Daiei Studios, know…"

"0.18931" "This is a beautiful, but simple film about…"

"-0.66898" "If you loved the 1993 movie (erotic, scien…"

"0.4883" "I am a great ex-male goalie, ex-rugby play…"

"0.48221" "This brief topic brought praise from all k…"

"0.39937" "This movie is really special! he is a youn…"

"0.78373" "I would give this movie high marks for fil…"

"0.34803" "I remembered seeing this program when I wa…"

"0.093937" "First, ignore the comment about how South …"

"-0.008254" "I do not see what the problem is with SOME…"

"0.44542" "William Faulkner was one of the American w…"

"**0.93213**" "Like many of today's movies are guilty, th…"

"-0.037253" "his is without a doubt the most exciting a…"

"0.5303" "This is not your typical Indian movie. The…"

"0.12745" "SPOILERS FORWARD During the first ten minu…"

"0.27931" "When I see "Plots with a View" (called "Un…"

The output sentiments appear to be reflective of reality, though there are some key points to note. While a few reviews have very strong positive or negative opinions, many of the opinions are neutral. While this may be accurate in terms of NLP, the sentiment approximation from human judgement might be different than the output values from the classifier. At least two instances, noted in red, appear to be negative reviews misclassified as positive reviews.

In future work, I would like to expand the languages surveyed to include non-Latin languages such as Japanese and Chinese. These languages have some key differences compared to languages like English and Spanish. These languages have a reputation for being difficult to machine-translate when compared to Romance languages. Concepts in these languages are represented in pictographs and sentence structure can differ from English (subject-object-verb vs. subject-verb-object). Testing these languages would probably require a word embedding and opinion lexicon designed for Japanese/Chinese text.

**References**

[1] <https://monkeylearn.com/sentiment-analysis/>

[2] <https://www.mathworks.com/matlabcentral/fileexchange/66229-text-analytics-toolbox-model-for-fasttext-english-16-billion-token-word-embedding>

[3] <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>