Sentiment Analysis on Movie Review Data Using Machine Learning Approach

Team Binary Beast
Mohee Datta Gupta, Rohith R, Sridhar M

Analysis procedure - Text preprocessing

- Dataset of movie reviews created using the source - github
- 2000 movie review posts - with equal positive and negative reviews

Text preprocessing:

- URL removal
- Bracket and number removal
- Tokenization
- Omitting punctuation
- Case conversion
- Omitting stop words
- Stemming

Analysis procedure

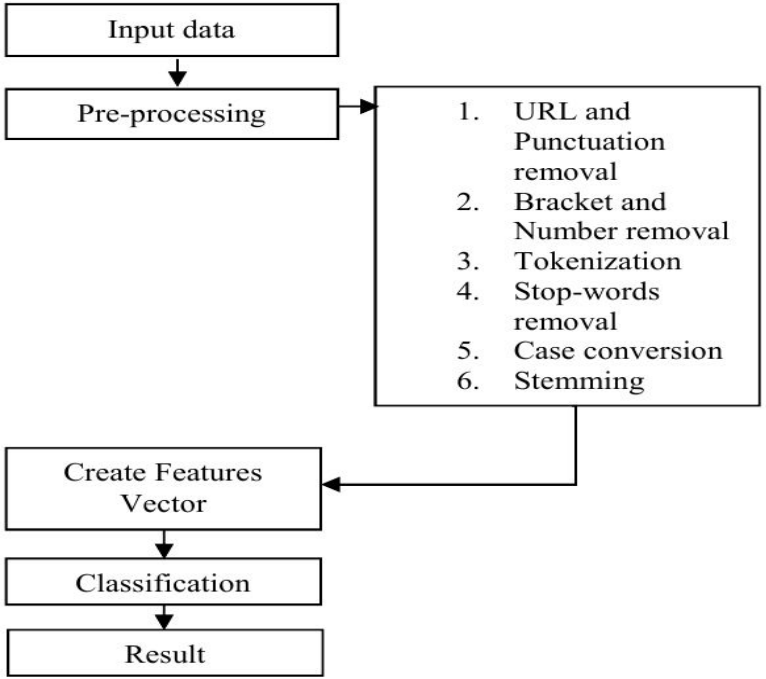

Fig. 1: Steps for classification in SA

Analysis procedure - Feature Vector Creation

- Classify each review as a positive or negative
- Every review is considered as a document
- Parts of Speech (PoS) tags are added to get more accurate sentiments
- Positive and negative sentiments are maintained by respective sentiment scores
- Every word has a feature
- No. of features in the document gives us its dimmensionality

Considered Classifiers

1. Naive Bayes classifier
    a. Bernoulli NB classifier
    b. Multinomial NB classifier
2. SVM
3. Maximum entropy classifier
4. Decision Tree classifier

Dataset

	URL	Text	Sentiment
2	http://www.imdb.com/title/tt0210075/usercr	Girlfight follows a project dwelling New York high school girl from a sense of futility into the world of amateur boxing where she finds self	POS
3	http://www.imdb.com/title/tt0337640/usercr	Hollywood North is an euphemism from the movie industry as they went to Canada to make movies because of tax breaks and cheaper cos	POS
4	http://www.imdb.com/title/tt0303549/usercr	That '70s Show is definitely the funniest show currently on TV. I started watching it about two and a half years ago, and as soon as I saw it	POS
5	http://www.imdb.com/title/tt0716825/usercr	9/10- 30 minutes of pure holiday terror. Okay, so it's not that scary. But it sure is fun.The Crypt Keeper (John Kassir) tales a tale of holi	POS
6	http://www.imdb.com/title/tt0182225/usercr	A series of random, seemingly insignificant thefts at her sister's boarding house has Miss Lemon quite agitated. A ring, light bulbs, a rucksa	POS
7	http://www.imdb.com/title/tt0347779/usercr	A very good adaptation of the novel by amrita pritam. Urmila and manoj bajpai have given their best.there is a natural flair in the movie an	POS
8	http://www.imdb.com/title/tt0095655/usercr	Ah, Moonwalker, I'm a huge Michael Jackson fan, I grew up with his music, Thriller was actually the first music video I ever saw apparently. B	POS
9	http://www.imdb.com/title/tt0298131/usercr	Although the beginning of the movie in New York takes too long, the movie is a must see for people who like this genre. When Hannah goe	POS
10	http://www.imdb.com/title/tt0088915/usercr	As many reviewers here have noted, the film version differs quite a bit from the stage version of the story. I have never seen the stage vers	POS
11	http://www.imdb.com/title/tt0828154/usercr	Bear in mind, any film (let alone documentary) which asserts any kind of truth, will generate an adverse and proportional amount of cynicis	POS

- The dataset has 3 columns : URL, Text, Sentiment
- The URL is useless for our project and hence dropped immediately
- The Text column has the movie review and will be pre-processed before using as the X_train
- The Sentiment column stores POS for a positive review and NEG for a negative review which is converted to a binary before using as our Y_train.

Contributions Plan (Tentative)

Mohee Datta Gupta - Data Analysis, Naive Bayes, MNB

Rohith R - Decision Tree, SVM

Sridhar M - SVM, ME classifier

Tentative Weekly Work Execution Plan

Week	**1 (4th-10th Apr)**	**2 (11th-17th Apr)**	**3 (18th-23rd Apr)**
Mohee	Figure out data pre-processing	Implementation of Naive Bayes & MNB	Concluding from results
Rohith	Implementation of Dec. Tree	Implementation of SVM & Dec. Tree	Comparing results from different methods
Sridhar	Implement Maximum entropy classifier & SVM	Implementation of SVM	Comparing results from different methods

Thank You!!