Sentiment Analysis

Bala Subrahmanyam Varanasi
Twitter: @vabasu

Agenda

- Quick Overveiw
  - Introduction
  - Motivation
  - Application Areas
  - Challenges
- Sentiment Analyzer for Telugu language.
  - A General Model
  - Our Approach
  - Future Work

“What people think?”

What others think has always been an important piece of information

“Which mobile should I buy?”

“Which schools should I apply to?”

“Which Professor to work for?”

“Whom should I vote for?”

“So whom shall I ask?”

Pre Web

- Friends and relatives
- Acquaintances
- Consumer Reports

Post Web

“…I don’t know who..but apparently it’s a good phone. It has good battery life and…”

- Blogs (google blogs, livejournal)
- E-commerce sites (amazon, ebay)
- Review sites (CNET, PC Magazine)
- Discussion forums (*forums.craigslist.org, forums.macrumors.com*)
- Friends and Relatives (occasionally)

“Whoala! I have the reviews I need”

Now that I have “too much” information on one topic…I could easily form my opinion and make decisions…

Is this true?

Who is going to read that?

…Not Quite

- Searching for reviews may be difficult
  - Can you search for opinions as conveniently as general Web search?
    eg: is it easy to search for *"iPhone vs Google Phone"*

- Overwhelming amounts of information on one topic
  - Difficult to analyze each and every review
  - Reviews are expressed in different ways

"the google phone is a disappointment…."

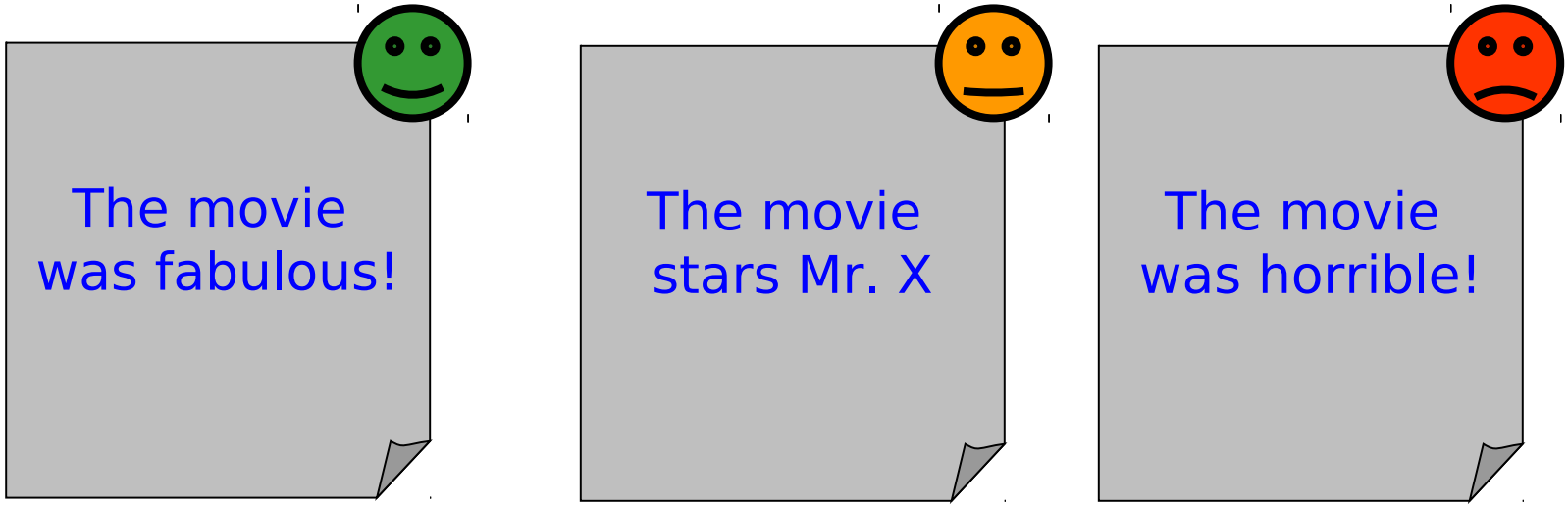
"don't waste your money on the g-phone…."

"google phone is great but I expected more in terms of…"

"…bought google phone thinking that it would be useful but…"

Introduction

- Sentiment Analysis
  - Determines the Attitude/Opinion/Sentiment of text by an author.
  - Aka - Opinion Mining
  - Uses NLP and CL to automate the extraction or classification sentiment.



Introduction

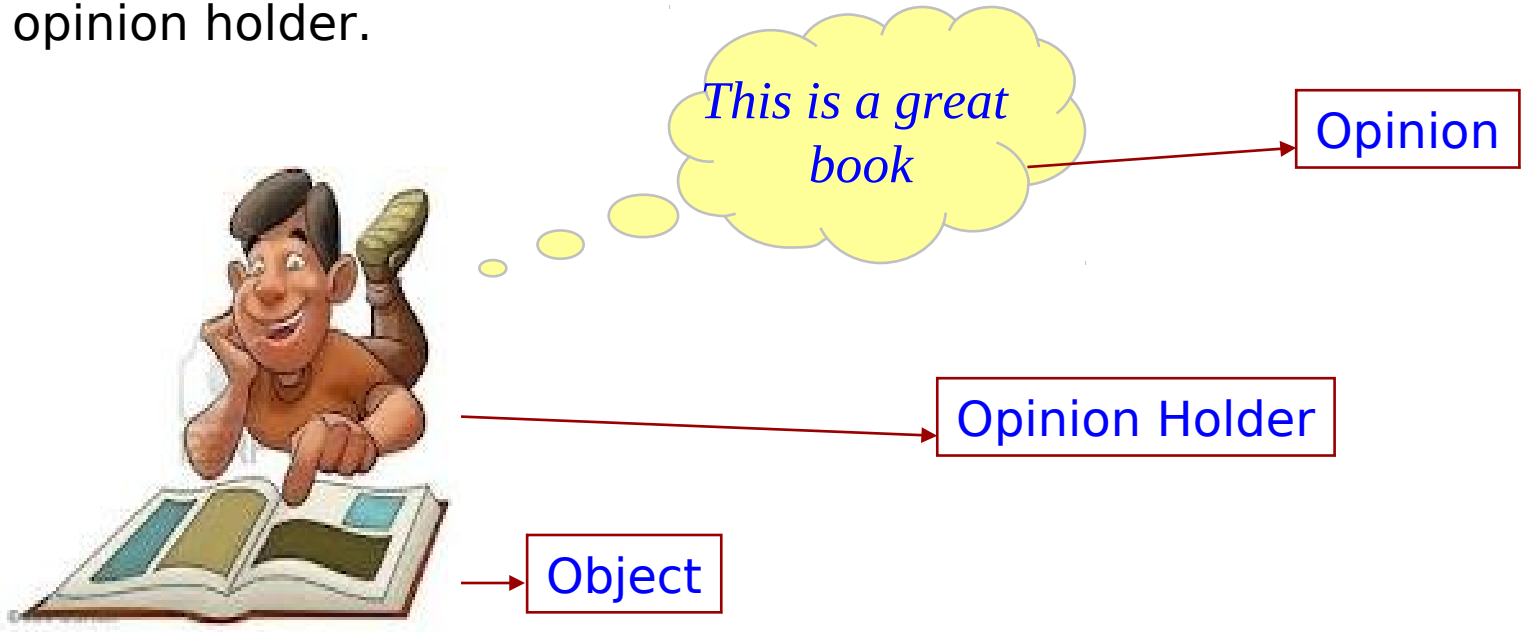
- Textual Information is categorized into two types.
  - Facts and
  - Opinions
- Facts are objective expressions about entities, events and their properties.
- Opinions are usually subjective that describe people's sentiments

Terms

- Sentiment
  - A thought, view, or attitude especially one based mainly on emotion instead of reason.
- Sentiment Analysis
  - aka opinion mining
  - use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text

Some basics...

- Basic components of an opinion
  1. **Opinion holder**: The person or organization that holds a specific opinion on a particular object
  2. **Object**: item on which an opinion is expressed
  3. **Opinion**: a view, attitude, or appraisal on an object from an opinion holder.



Motivation

- Businesses and organizations:
  - Product and service benchmarking.
  - Market intelligence.
- People:
  - Finding opinions while purchasing a new product.
  - Finding opinions on political topics.
- Advertisement:
  - Placing ads in the user-generated content.
  - Place an ad when one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- Information search & Retrieval:
  - Providing general search for "opinions".

Challenges

- Sentiment and Subjectivity Classification
- Feature based Sentiment Analysis.
- Sentiment analysis of Comparitive Sentences
- Opinion search and retrieval
- Opinion span and utility of opinions.

Applications

- Applications to review-related websites
- Applications as a sub-componenet Technology
- Applications in business and government Intelligence
- Applications across different domains

Sentiment Analyzer for Telugu

Using Telugu Movie Review as Corpus.

Sentiment Analysis Model

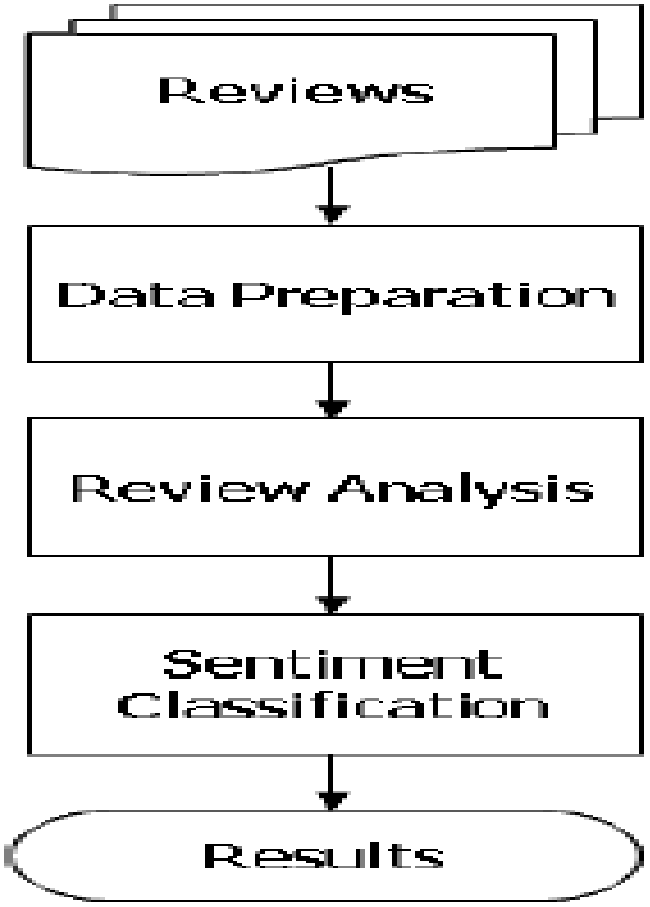


Figure 1: A typical sentiment analysis model.

Data Preparation

- It performs data preprocessing and cleaning on the dataset.
  - Eg: Removing non-textual contents and markup tags (for HTML pages)

- Balance training datasets distributions.

Review Analysis

- In this, the linguistic features of reviews like opinions and/or product features, can be identified.

- Two commonly adopted tasks for review analysis
  - POS tagging
  - Negation tagging.

Sentiment Classification

- There are two main techniques for sentiment classification:
  - The Symbolic technique uses manually created rules and lexicons.
  - The Machine Learning approach uses Supervised or Un-Superviced Learning to construct a model from a large training corpus.

Methodology

- Our method of sentiment analysis is based upon machine Learning.
- It uses...
  - Large set of Telugu Movie Reviews as Corpus.
    - It containg above 106 movie reviews as our data set. And it is classified by subjectivity/Objectivity and negative/positive attitude, manually.
  - Bag-of-words model to extract Text features.
  - Supervised Learning algorithm – Naive Bayes
  - NLTK for implementing these algorithms.

Machine Learning Implementation

- Polarity detection
  - 106 positive & 106 negative movie reviews from telugu.oneindia.com.
  - Preprocessing of data:
    - Tokenizing
    - Stop word removal
  - Feature set definition using frequency distribution.
  - Training the classifier using 'Naïve Bayes’.
  - Applying classifier to find the polarity of the reviews

Pre-processing stage

- Examples for Tokenizing text.
  - Text into sentences

    In [1]: import nltk

    In [2]: from nltk.tokenize import sent_tokenize

    In [3]: para = "వినాయక్ దర్శకత్వంలో రూపొందిన ఈ చిత్రంలో అల్లు అర్జున్ కు జంటగా తమన్నా హీరోయిన్ గా నటించింది. తొలిసారి వినాయక్ సినిమాకు కీరవాణి బాణిలను అందించారు."

    In [4]: sent_tokenize(para)
  - Sentences into words

    In [5]: sentence = "ఎయిట్ పాక్స్ గల సంజయ్ సింగానియా(అమీర్ ఖాన్) సెల్ ఫోన్స్ రంగంలో పేరున్న పెద్ద పారిశ్రామిక వేత్త."

    In [6]: from nltk.tokenize import SpaceTokenizer

    In [7]: tokenizer = SpaceTokenizer()

    In [8]: tokenizer.tokenize(sentence)

Filtering Stopwords

- Code snippet to perform filtering operation.

In [1]: from nltk.corpus import stopwords

In [2]: telugu_stopwords = set(stopwords.words('telugu'))

In [3]: words = ["రాజమౌళి","మరో సారి", "తన","దర్శకత్వ","ప్రతిభను","నిరూపించుకొన్నార"]

In [4]: [word for word in words if word not in telugu_stopwords]

Out[5]: […]

Bag of words Model

- It takes individual words in a sentence as features, assuming their conditional Independence.
- Bag of words is a model that takes

  In [5]: words = ["రాజమౌళి","మరో సారి", "తన","దర్శకత్వ","ప్రతిభను","నిరూపించుకొన్నార"]

  In [6]: def bag_of_words(words):

  ....: return dict([(word, True) for word in words])

  In [7]: bag_of_words(words)

  Out[8]: {'\xe0\xb0\xa4\xe0\xb0\xa8': True, ... }

  - We represent the feature vector as a python dictionary; NLTK, for example, uses this representation as shown above.

Naive Bayes Classifier

- NaiveBayesClassifier, uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label

  .

- The formula is:

  P(label | features) = P(label) * P(features | label) / P(features)

  where

  - P(label) is the prior probability of the label occurring
  - P(features | label) is the prior probability of a given feature set beingclassified as that label.
  - P(features) is the prior probability of a given feature set occurring.
  - P(label | features) tells us the probability that the given features should have that label.

DEMO TIME

Future Work

- Developing WordNet for Telugu Langauge.
- Handling Syntactic and Semantic properties
- Handling Negation.

References

- Natural language toolkit. http://www.nltk.org/
- S. Bird, E. Klein, and E. Loper. Natural Language Processing with Python – Analyzing Text with NLTK. O'Reilly Media, 2009.
- B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. Now Publishers Inc, July 2008.
- Report on "Sentiment Analysis of Movie Review Comments" by Kuat Yessenov, Sa˘ a Misailovi .
- Telugu Movie Reviews from http://telugu.oneindia.in/.
- Opinion Mining Short Tutorial by Kavitha Ganesan and Hyun Duk Kim