Sentiment Analysis

-Aditya Yadavalli

Paper 1 Introduction

- Paper is on Sentiment Analysis on Online Product Review
- Done by Raheesa Safrin, K.R Sharmila, T.S Shri Subangi, E.A Vimal
- Sentiment is an emotion or an attitude prompted by the feelings of the customer. Sentiment analysis studies people’s opinion towards the product
- Challenges:
    - Fake Comments
    - One situation positive other situation negative.

Paper 1 Proposed System

- Objective - review the product based on the comments given by the customers
- These comments are classified into positive & negative
- How ? Steps:
  - Tokenise
  - POS tag
  - Negation Phrase Identification algorithm

Paper 1 Implementation

- Steps:
  - Data collection
  - Pre-processing (NLP)
  - Feature Labelling
- Data Collection
  - Data is collected in two forms:
    - Star rating
      - Count of the star is mapped to certain adjectives
    - Textual (Emojis not being considered)

Paper 1 Implementation (1)

- Pre-processing
  - Steps:
    - POS tagging
    - Negative Phrase Detection Algorithm
  - Found:
    - Adjectives and adverbs express more polarity
  - Also does:
    - Remove/replace noisy and incomplete data

Paper 1 Implementation (2)

- Feature Extraction:
  - Steps:
    - The input is converted into features (feature vectors)
    - For feature labelling they use two files - positive, negative
    - The positive words are labeled as ‘0’ whereas negative are labeled under ‘1’.
  - K-means Cluster:
    - Clustering used to classify the retrieved dataset through a number of clusters
    - The labelled words are now taken for clustering
    - We get positive, negative clusters

Paper 1 Performance Analysis

- Recall
  - Recall is the proportion of real positive cases that are correctly predicted positive.
  - Recall = tp/(tp+fn)
- Precision
  - It is a proportion of predicted positive cases that are correctly real positive.
  - Precision = tp / (tp+fp)
- Accuracy
  - Accuracy represents what percent of prediction were correct
  - Accuracy = (tp+tn)/ (tp+tn+fp+fn)
- This test has:
  - Recall - 90%
  - Precision - 87%
  - Accuracy - 90.47%

Paper 2 Introduction

- Paper is on Sentiment Analysis on Twitter
- Done by Akshi Kumar and Teeja Mary Sebastian
- This research carried out by them was done to use sentiment analysis to gauge the public mood and detect any rising antagonistic or negative feeling on social medias
- Data characteristics:
  - Message length : 140 characters max
  - Writing technique: use of acronyms, cyber slang, misspell and emoticons.
  - Availability: Lot of data is available on Twitter
  - Topics: Very diverse
  - Real Time: Tweets on the other hand being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events.

Paper 2 System Architecture

- Pre-processing of Tweets:
  - Remove all URLs (e.g. www.example.com), hash tags (e.g. #topic), targets (@username), special Twitter words (“e.g. RT”).
  - Calculate the percentage of the tweet in Caps.
  - Correct spelling. A sequence of repeated letters is tagged by its weight. They do this to differentiate between emphasised usage and regular usage.
  - Replace all the emoticons with their sentiment polarity
  - Remove all punctuation marks after counting the no. of exclamation marks

Paper 2 System Architecture (1)

- Scoring module (mainly depends on adj, adv and verbs)
  - Adjectives:
    - Corpus based approach is used cause adjectives are usually domain specific
    - we ascribe same semantic orientation to conjoined adjectives in most cases and in special cases when the connective is “but”, the situation is reversed.
  - Adverbs and Verbs:
    - Use dictionary methods since they are not domain dependent
    - Use seed list and then increase your dictionary by finding antonyms and synonyms.
    - Antonyms have opp semantic orientation and synonyms the same

System Architecture (2)

- Scoring Module (continued)
  - Tweet scoring
    - |OI(R)| denotes size of opinion set extracted
    - Pc denotes fraction of tweet in caps
    - Ns denotes the count of repeated letters
    - Nx denotes the count of exclamation marks
    - S (AGi) denotes score of the ith adjective group
    - S (VGi) denotes the score of the ith verb group
    - S (Ei) denotes the score of the ith emoticon
    - Nei denotes the count of the ith emoticon.

$$S(T) = \frac{(1 + (P_c + \log(Ns) + \log(N_x))/3)}{|OI(R)|} * \sum_{i=1}^{|OI(R)|} S(AG_i) + S(VG_i) + N_{ei} * S(E_i)$$

Paper 3 Introduction

- Paper is titled as Twitter Sentiment Analysis: The Good the bad and the OMG!
- Research done by Efthymios Kouloumpis, Theresa Wilson, Johanna Moore
- Challenges
  - Will message length constraints change language ? How to take care of that ?
  - Breadth of the topics covered

Paper 3 Data

- There are 3 different corpora of Twitter messages in their experiments
- For development and training:
  - Hashtagged data set (HASH) by Edinburgh Twitter corpus
  - Emoticon data set (EMOT) by http://twittersentiment
- For evaluating:
  - Manually annotated data set produced by the iSeive Corp.

Paper 3 Data (1)

- Hashtagged data set:
  - Filter out:
    - Duplicate tweets
    - non-English tweets
    - Tweets that don’t have hashtags
- Emoticons data set:
  - The Emoticon data set was created by Go, Bhayani, and Huang for a project at Stanford University
  - Messages containing both positive and negative emoticons were filtered out
- iSeive:
  - Manually tagged
  - Used for evaluation *only*

Paper 3 Data (2)

- Preprocessing :
  - Tokenization
    - Abbreviations (BRB, OMG) are identified as one token.
  - Normalisation
    - Abbreviations are replaced by their actual meaning
    - Noting caps that emphasise the meaning and then lowering the case
    - Presence of #, @, and URLs are noted and replaced with placeholders
  - POS tag
    - Normalisation improves POS tagging

Paper 3 Features

- N-gram features
  - Remove stop words
  - Negation detection
  - Unigrams and bigrams are identified and ranked acc to info gain
- Lexicon Features
  - Directly use MPQA subjectivity lexicon
- POS features
  - Features for number of verbs, nouns etc in each tweet
- Micro-Blogging Features
  - Emoticons
  - Abbreviations
  - Intensifiers (All caps, repetitions)
  - Use Internet Lingo Dictionary

Paper 3 Result

- Interestingly, the best performance on the evaluation data comes from using the n-grams together with the lexicon features and the microblogging features. Including the part- of-speech features actually gives a drop in performance.