Project

Twitter Profile Building by Sentiment Analysis

Project Guide: Dr. Mary Saira Bhanu

Members:

Lenoy Jacob	106108041
Ashok Kumar	106108052
Anirudh S	106108053

Objectives

- Micro-blogs are a challenging new source of information for data mining techniques.
- They are social tools and therefore, are a source of valuable real-time data.
- Twitter is a very popular microblogging service where users create status messages (called "tweets"), which reflect the personality of the user.
- Analyzing these tweets can help us build a personality profile of the user, which would help other users follow the people they want.

Procedure

- Data Collection
- Keyword Classification
- Sentiment Analysis
- Association of sentiments with keywords
- Building the profile of the user
- e.g. If a user tweets positively about Android and negatively about iPhone, output would be:

  <username> (Android +) (iPhone -)

Steps Involved

- Sentiment Analysis
- Binary SVM classifier
- OpenNLP POS tagger
- Word stemming
- Syntactic parsing
- Handling Target-independent Features
- Handling Extended Targets
- Handling Target dependent features
- Building a corpus by searching for keywords from a white list
- Evaluation and representation

Approach

- Subjectivity classification
- Polarity classification

- In each of the steps, a binary SVM classifier is built to perform the classification.

Preprocessing

- Tweet Normalization
- POS tagging
- Word Stemming
- Syntactic Parsing

Other kinds of Tweets

- Retweets
- Tweets containing the target and published by the same person.
- Tweets replying to or replied by the tweet are to be classified, as well.

Extended Targets

- People express their sentiments about a target by commenting not on the target itself but on some related things of the target.
- One may express a sentiment about a company by commenting on its products or technologies.
- Example:

  “I am passionate about Microsoft technologies especially Silverlight.”

Target Dependent Classification

- Syntactic Parse Tree
  - Verb
  - Noun
  - Adjectives
  - Adverbs
  - Negation
- “iPhone works better with the Cell-Band”, will generate the feature “arg1_v_well”

Classifying using SVM

- Special binary feature indicating whether or not the tweet contains at least one of the target-dependent features.
- If the feature is present, the entry will be 1; otherwise it will be 0.
- This will be done by using support vector machines.

Tweet Corpus

- No Target dependent Tweet Corpus available
- Build using Twitter search API
- Query using a list of keywords
- Manually classify each tweet as positive, negative or neutral towards the query
- Duplicate Tweets

Query Topics

- Sports
  - Cricket
  - Tennis
  - Football
- Celebrities
  - Sachin
  - Federer
  - Pele
- Sporting Events
  - World Cup
  - Grand Slams
  - Leagues

Accuracy

- Human accuracy – 80%
- Accuracy of Target Independent Classifying algorithms
- Accuracy of Target Dependent Classifying algorithms
- 10 fold cross validation

Results and Representation

- Accuracy of matching
  - Comparison tables
  - Confusion matrices

- Representation
  - <username>

    <keyword 1><sentiment 1>

    <keyword 2><sentiment 2>

    ...

Implementation

- Binary SVM classifier - *http://svmlight.joachims.org/*
- OpenNLP POS tagger - *http://opennlp.sourceforge.net/projects.html*
- Word stemming – word stem mapping table (or) snowball
- Syntactic parsing - MST dependency parser
- Target-independent Features – hashtags, emoticons and sentiment lexicon features (using General Inquirer)
- Extended Targets
- Target dependent features – adjectives, adverbs, noun, verb
- Building a corpus by searching for keywords
- Evaluation and representation
- Hosted at *https://github.com/anirudh24seven/B.Tech-project*

References

- Target-dependent Twitter Sentiment Classification - Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao (http://www.aclweb.org/anthology/P/P11/P11-1016.pdf)

  published in The Association for Computational Linguistics (ACL) journal
- Opinion Mining and Sentiment Analysis - Bo Pang and Lillian Lee

  (http://dl.acm.org/citation.cfm?id=1454712)

  published in ACM journal Volume 2 Issue 1-2, January 2008
- Sentiment Knowledge Discovery in Twitter Streaming Data - Albert Bifet and Eibe Frank (http://www.cs.waikato.ac.nz/~eibe/pubs/Twitter-crc.pdf)
- Twitter Sentiment Analysis - Alec Go, Lei Huang, Richa Bhayani (http://twittersentiment.appspot.com/)

Thank you