Project

Twitter Profile Building by Sentiment Analysis

The Team

Project Guide: Dr. Mary Saira Bhanu

Members:

Lenoy Jacob	106108041
Ashok Kumar	106108052
Anirudh S	106108053

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

1. Subjectivity classification to decide if the tweet is subjective or neutral about the target

2. Polarity classification to decide if the tweet is positive or negative about the target if it is classified as subjective in Step 1

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

Preprocessing

Rich feature representations are used to distinguish between sentiments expressed towards different targets.

In order to generate such features, the following have to be done:

- Tweet normalization

  Correct simple spelling errors and variations

- POS tagging

  Implement Part of Speech tagging

- Word stemming

  Stem words by using a word stem mapping table

- Syntactic parsing

  Parse words by using a MST dependency parser

Extended Targets

It is quite common that people express their sentiments about a target by commenting not on the target itself but on some related things of the target.

For example, one may express a sentiment about a company by commenting on its products or technologies.

It is assumed that readers or audiences can clearly infer the sentiment about the target based on those sentiments about the related things.

For example, in the tweet
“I am passionate about Microsoft technologies especially Silverlight.”
the author expresses a positive sentiment about “Microsoft” by expressing a positive Sentiment directly about “Microsoft technologies”.

Target-dependent classification

Target-dependent sentiment classification needs to distinguish the expressions describing the target from other expressions. Syntactic parse trees can be used for this.

In addition to the noun phrases including the target, we further expand the extended target set with the following three methods:

1. Adding mentions co-referring to the target as new extended targets. It is common that people use definite or demonstrative noun phrases or pronouns referring to the target in a tweet and express sentiments directly on them.

2. Identifying the top K nouns and noun phrases which have the strongest association with the target. Here, we use Pointwise Mutual Information (PMI) to measure the Association.

3. Extracting head nouns of all extended targets, whose PMI values with the target are above some predefined threshold, as new extended targets.

Target-dependent Features

Syntactic parse tree

•Verb
•Noun
•Adjective
•Adverb

For example, for the target iPhone in the tweet “iPhone works better with the Cell-Band”, we will generate the feature “arg1_v_well”

Moreover, if any word included in the generated target-dependent features is modified by a negation, then we will add a prefix “neg-” to it in the generated features.

Classifying using SVM

To overcome the sparsity of target-dependent features mentioned above, we design a special binary feature indicating whether or not the tweet contains at least one of the above target-dependent features.

Target-dependent features are binary features, each of which corresponds to the presence of the feature in the tweet. If the feature is present, the entry will be 1; otherwise it will be 0.

This will be done by using support vector machines.

Other kinds of tweets

The following three kinds of related tweets are also used as context for a tweet.

1. Retweets. Retweeting in Twitter is essentially the forwarding of a previous message. People usually do not change the content of the original tweet when retweeting. So retweets usually have the same sentiment as the original tweets.

2. Tweets containing the target and published by the same person. Intuitively, the tweets published by the same person within a short timeframe should have a consistent sentiment about the same target.

3. Tweets replying to or replied by the tweet are to be classified, as well.

Tweet Corpus

Because there is no annotated tweet corpus publicly available for evaluation of target-dependent Twitter sentiment classification, we have to create our own.

For each of those queries, we are planning to download tweets containing the query using the Twitter API.

We manually classify each tweet as positive, negative or neutral towards the query with which it is downloaded.

Duplicate tweets have to be removed as well.

Query topics

Sports:

The acual sports:
Cricket
Football
Tennis

Sports celebrities:
Sachin
Pele
Federer

Sporting Events:
Olympics
World Cup
Leagues
Grand Slam opens

Experiments and accuracy

We will conduct several experiments to evaluate subjectivity classifiers using different features. In the experiments, we will consider the positive and negative tweets annotated by humans as subjective tweets.

Human accuracy - 80%

10 fold cross validation

Break data into 10 sets of size n/10.
Train on 9 datasets and test on 1.
Repeat 10 times and take a mean accuracy.

Results and representation

Results to be shown:

Important keywords matching with the person
Association of the sentiment with those keywords

Representation:

Confusion matrices to represent accuracy

<username>
(<keyword 1> <association 1>, <keyword 2> <association 2>...)

Tools used

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 - Maximum Spanning Tree dependency parser

Target-independent Features – hashtags, emoticons and sentiment lexicon features (using General Inquirer http://www.wjh.harvard.edu/~inquirer/)

Extended Targets

Target dependent features – adjectives, adverbs, noun, verb

Building a corpus by searching for keywords related to sports

Evaluation and representation – Accuracy tables, Confusion matrices