SENTIMENT ANALYSIS - INTRODUCTION

Author:

Patrick Schneider

Course:

MIRI – Algorithms for Data Mining

Date:

26.05.2017

University:

UPC FIB Barcelona, Spain

ABSTRACT

In every aspect of life, part of our information-gathering behavior has been to find out what other people think. With the growing accessibility of opinion sources such as personal blogs, social networks and review platforms, new opportunities and challenges emerge. Sentiment analysis deals with the computational treatment of opinion, subjectivity and sentiment in text.

This essay gives an overview in the fields of sentiment analysis and covers techniques and approaches that must be considered after determining the direction and goal of an project.

"Your most unhappy customers are your greatest source of learning." Bill Gates

Table of Content

1. INTRODUCTION – SENTIMENT ANALYSIS

Over the last years, sentiment analysis (also known as opinion mining) has won more and more interest. The basic idea is to automatically classify a text written into a positive or negative feeling. In some cases, it is even for humans difficult to categorize an opinion to a specific feeling. The interpretation is influenced by e.g. cultural factors, individual experiences, bad writing, missing background information, etc.. Short text, like it is written on social networks, make the task even more difficult.

2. CHALLENGE

Individuals express opinions in complex ways, which makes understanding the subject of human sentiments a hard issue to tackle. Rhetorical devices like sarcasm, irony, and implied meaning can mislead sentiment analysis, which is why concise and focused opinions like product, book, movie, and music reviews are easier to analyze.

There are difficulties in the sentiment analysis and the evaluation process. These difficulties aggravate the analyzation and affect the accuracy of sentiments and their polarity. Like earlier mentioned, those are the major challenges [1]:

- Emoticon detection
- Spam and fake review detection
- Negation and BI-polar words
- Sarcasm
- Neutral opinion
- Domain Dependency
- NLP over heads

3. APPLICATION FIELDS

Sentiment analysis is frequently used as a part of business intelligence to understand the subjective reasons why customers are or are not reacting to something. Another application field that has developed over the recent years are political analysis. Political parties use sentiment analysis for voter analytics. Further application fields can be found in sociology and trend analysis. Following initial questions can be analyzed:

- Why are customers purchasing an item?
- What do they think about the user experience?
- Did customer service support meet their expectations?
- What is the current opinion of person X?
- How did people react to event x?

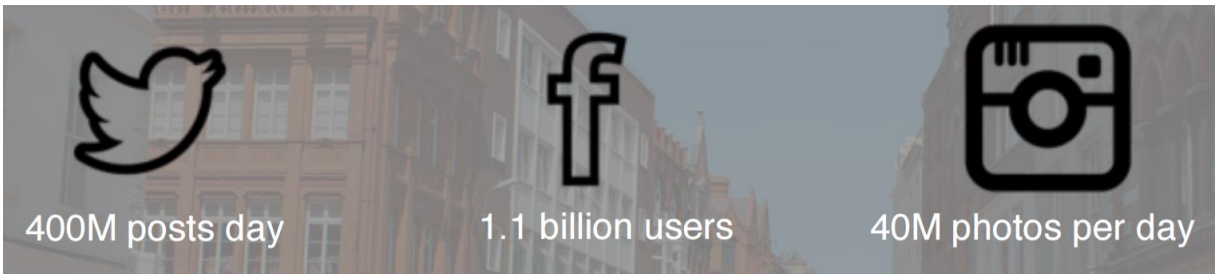
Sentiment analysis can likewise be utilized as a part of the research fields of political science, human science, and brain research to dissect patterns, ideological predisposition, gage responses, and so on.

4. METHOLOGIE

A. Data Collection

In the learning based approach, concerning the data, it is important to have a domain specific corpus of documents and texts. This data can be used to train the algorithm of the later feature selection and prediction. This data can be already classified or needs to be classified in the first step. There are already existing databases with domain specific labeled entries.

The actual prediction field is huge. Social media offers a big environment for prediction.



Other platforms that offers people to give an opinion can be used too. Example for this are Amazon Reviews, IMDB Reviews.

B. Pre-Processing and Noise Removal

Tokenization

Tokenization is the way toward separating a flood of text into words, expressions, symbols, or other significant components called tokens. The input of tokens is the basis for further processing, like parsing or message mining.

Part-of-speech (POS) tagging

Part-of-speech tagging is the assignment of words and punctuation marks of a text to part of speech. For this purpose, the definition of the word as well as the context (e.g. adjoining adjectives or nouns) is taken into account.

Stop words

Stop words are frequently-used words, that have little value in helping to select matching a user needs. Those words are excluded from the vocabulary entirely. In some contexts, they have a different impact. Stop words are generally determined by the frequency they appear.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Figure 1: 25 stop words

Lemmatization

In lemmatization, the words lemma is searched and based on the corpus context changed to a unifying word. This is archived with the understanding of the context, the lexicon and part-of-speech tagging.

Example, “better” is related to “good”, “running” is related to “walk” and so on.

Stemming

In Stemming a word gets reduced to its root form(stem). The root form must not be a word. The idea is to generate the actual word by concatenating the right suffix.

Example, the words fish, fishes and fishing stem into fish - that is a right English word. On the other side, the words study, studies and studying stems into studi - which is not an English word.

Additional improvements

Further improvements can be done by removing unnecessary information and symbols like:

- Numbers
- URLs/Links
- Punctuation
- Whitespace
- Etc.

C. Domain classification

Domain classification means to classify the data to specific subject domains like markets, economy, industry, technology, presidential elections, etc. Depending on the domain, different features are important for the algorithms. Each domain should have its own classifier.

Example: A positive news for Apple may be a negative news for Microsoft stocks. Positive news about the dollar currency, may have a negative impact on the crude oil value.

D. Named Entity Recognition

Named entity recognition (NER) is a major area of research in machine learning and natural language processing. NER is used to answer many real-world questions:

- Does a tweet contain the name of a person? Does the tweet also provide his current location?
- Which companies were mentioned in a news article?
- Were specified products mentioned in complaints?

There are mainly two kind of approaches for building NER:

- Rule Based: This approach uses heuristics to determine sentiments. It uses linguistics and communications research to analyze sentiments.
- Statistics or Machine learning based: A data driven approach which uses the labeled corpus of texts and their sentiments to predict.

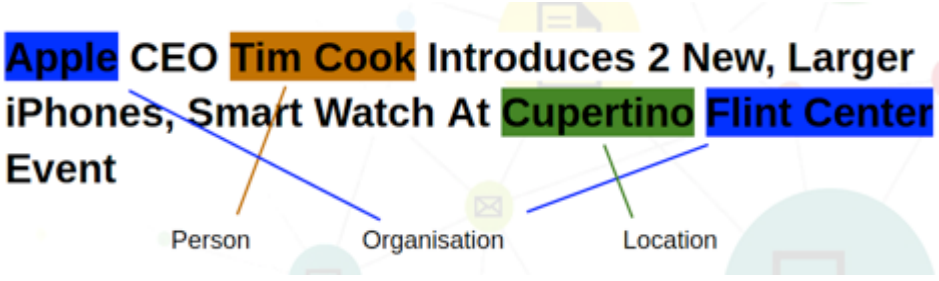


Figure 2: Example of name entity

E. Subjectivity Classification

Sentences must be categorized in to subjective or objective, since subjective sentences hold sentiments and objective sentences are facts and figures.

Example:

- The battery life of this camera is very good.
- Camera is a good device for capturing photographs.

Both sentences contain opinion with the word “good”. The first sentence is subjective and second one is objective. The goal of subjectivity classification is to restrict unwanted and unnecessary texts from processing. However, classifying a sentence as either subjective or objective is a complex task, because there is very little availability of training dataset. Annotated sets of subjective and objective sentences are difficult to obtain and requires lots of manual processing. [3]

F. Feature Selection

In text classification, the feature selection is the process of selecting a specific subset of the terms of the training set and using only them in the classification algorithm for sentiment extraction. The feature selection process takes place before the training of the classifier.

There are different feature selection approaches:

- frequency-based
- part-of-speech based
- lexicon selection based

Frequency based selection

An often practice in text modeling is to remove words which appear rarely in the corpus. These are probably misspellings that do not help in generalization during classification. On the other hand, words that occur only once in the corpus have been found to be high-precision indicators of subjectivity.

Part of speech based selection

Part-of-speech information is commonly used in sentiment analysis. One simple reason is that part-of-speech tagging can be a crude form of word sense disambiguation. Subjectivity detection have revealed a high correlation between the presence of adjectives and sentence. The focus lays on detecting document sentiment based on selected phrases, where the phrases are chosen with a

number of pre-specified part-of-speech patterns, with an including adjective or adverb.

lexicon based selection

Sentiment-annotated lexicons can be used for feature selection. By selecting terms which are indicative of strong sentiment, less useful features may be excluded from the feature set. Popular lexicons are the extensions of WordNet a large lexical database of English. SentiWordNet, for example, contains polarity and objectivity labels for the WordNet terms

Negation support feature

The negation handling is an important task in opinion and sentiment related analysis. While in the feature representations of “I like this product” and “I don’t like this product” are very similar by most similarity measures, the only differing token is the negation term. The term gets classified in the opposite class. This situation has in opinion mining a bigger weight than in most other information retrieval algorithms. In sentiment analysis, the feature vector should contain a second order selection for text segments, that deals with negations.

Feature generalization

Example could be, for each sentence in the corpus, each entity name can be replaced by PARTY or OTHER. Patterns such as “PARTY will win,” “go PARTY again,” and “OTHER will win” can be used for better extraction. This scheme especially performs well using simple n-gram features, when classifying which party a given message predicts to win.

G. Sentiment Extraction

After the features are selected, they can be enriched with already existing feature lexicons. Now the models can now be trained with the test data. To extract the sentiments, unsupervised learning and supervised learning can be used. In the following are the five most known algorithms for this task described:

1. **Naive Bayes Classifie**r uses far less computing power compared to other methods and often is a baseline method for many models.

2. **Maximum Entropy Classifier** is a parameterized method and works by extracting features from the text and combining the features in a linear fashion for classification. This is a member of the log-linear or exponential family of classifiers.

3. **Decision Trees** works by creating a decision tree of root, branches and leaves, creating a decision point at every branch. The decision is taken at the leaf node

4. **TiMBL** is a classifier based on k-nearest neighborhood algorithm. This works based on the principle of solving a problem by leveraging the learnings of previously solved problems.

5. **Support Vector Machines** are treated as classifiers with high accuracy but have their own challenges to deal with skewed nature of datasets and computational complexity.

5. CONCLUSION

By analysis the topic domain of sentiment analysis, I learned that there are many connecting domains and it turned out to be more difficult to cover everything in a small introduction, than I initially expected to be able to. Over the course of the last four weeks I could get practical experience with the tokenization and name entity recognition process. A lot time researching was needed to cover sentiment analysis in a whole. The essay can be used as an initial discovery of the fields of sentiment analysis. While creating this writing, my interest was strengthened to discover analysis frameworks and practice twitter sentiments and visualizations. The business understanding, the mythology and techniques are covered with this essay and further research can be made by selecting the goals of an analysis and select the important elements, that were discussed in this writing.

6. REFERENCES

[1] Doaa Mohey El-Din Mohamed Hussein, *A survey on sentiment analysis challenges*, Journal of King Saud University, 2016

[2] Hassan Saif, Miriam Fernandez, Yulan He, Harith Alani, *On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter,* Knowledge Media Institute, The Open University, UK, 2014

[3] Ahmad Kamal, *Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources*, Jamia Millia Islamia (A Central University), 2014

[4] Yelena Mejova1 and Padmini Srinivasan, *Exploring Feature Definition and Selection for Sentiment Classifiers,* 2014

[5] Bo Pang and Lillian Lee, *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008