Twitter Sentiment Analysis

Daniele Gerbaldo, Thieu Nguyen, Tim Röthig
daniele.gerbaldo.001@student.uni.lu,
thieu.nguyen.001@student.uni.lu,
tim.rothig.001@student.uni.lu

University of Luxembourg,
Faculty of Science, Technology and Communication

December 2019

Abstract

Sentiment analysis refers to the use of machine learning, natural language processing, and computational linguistics to determine the sentiment content from the written language, i.e. it analyzes people's opinions, attitudes, and emotions for products, services, organizations, individuals, events or topics. In this project, we try to build the automatic sentiment analyzer system based on SVM and other techniques. The frequency of occurrence of words used as features for SVM. To evaluate our system, the Twitter dataset which containing tweets that are manually annotated for the sentiment (positive, negative or neutral) will be used. Eventually, we use the system to classify the unseen message's sentiments.

Keywords: Machine Learning, Sentiment Analysis, SVMs, Feature Selection, Twitter Dataset, Nature Language Processing

1. Introduction

Nowadays, the age of the Internet has changed the way people express their opinions and views. It is now mainly done through blog posts, online forums, product review websites, social media, etc. Millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their feelings, opinion and share views about their daily lives. Social media is generating a massive amount of sentiment data in the form of tweets, status updates, blog posts, comments, reviews, etc. Therefore, It is provided an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend on user-generated content online to a large extent to make decisions. For example, if someone wants to buy a product or wants to use any service, they firstly search for its reviews online, discuss it on social media before making a decision. The amount of content generated by users is too large for the normal user to analyze. Therefore, it is necessary to automate this, various sentiment analysis techniques are widely used. Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand their products or services in such a way that it can be offered as per the user's needs.

There are two types of machine learning techniques that are generally used for sentiment analysis, one is unsupervised and the other is supervised. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. Supervised learning is based on the labeled dataset and thus the labels are provided to the model during the process. These labeled datasets are trained to produce reasonable outputs when encountered during decision- making.

To help us to understand the sentiment analysis in a better way, this project is based on supervised machine learning. The rest of the paper is organized as follows. The second section discusses in brief about the work carried out for sentiment analysis in the different domain by various researchers. The third section is about the approach we followed for sentiment analysis. Section four is about implementation details and results followed by a conclusion and future work discussion in the last section.

2. Related Work

Work in the field "Sentiment analysis" has started since the beginning of the century. In its early stage, it was intended for binary classification, which assigns opinions or reviews to bipolar classes such as positive or negative. Paper [**?**] predicts review by the average semantic orientation of

a phrase that contains adjectives and adverbs thus calculating whether the phrase is positive or negative with the use of unsupervised learning algorithm which classifies it as thumbs up or thumbs down the review.

In [**?**] the product feature uses a latent semantic analysis (LSA) based filtering mechanism to identify opinion words that are used to select some sentences to become a rich review summarization. In another work [**?**], the polarity of the word is being calculated by all the words in the sentence, which can either be positive or negative depending on the related sentence structure. In [**?**] has proposed to pre-process the data to improve the quality structure of the raw sentence. They have applied the LSA technique and cosine similarity for sentiment analysis. [**?**] proposed a method based on verbs as an important opinion term for sentiment classification of a document belonging to the social domain. In [**?**] applied phrase pattern method for sentiment classification. It uses part of speech based rules and dependency relation for extracting contextual and syntactic information from the document.

Overall, a lot of work has also been done where researchers have explored and applied soft-computing approaches, mainly machine learning and neural works for sentiment analysis. In [**?**], the authors introduce a novel approach for automatically classifying the sentiment of Twitter messages. Their main idea is using tweets with emoticons for distant supervised learning. They apply several machine learning algorithms like Naive Bayes, Maximum Entropy, and SVM. In [**?**], the authors applied several ML techniques like Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature to analyze customer's review sentiment.

3. Methodology

A dataset is created using twitter (a popular microblogging service where users create status messages called tweets) posts of electronic products. These tweets sometimes express opinions about different topics. Therefore, they are short messages full of slang words and misspellings. So we propose a method to automatically classifying sentiment (positive, neutral or negative) from a tweet. This is done in three phases. In the first phase preprocessing is done. Then a feature vector is created using relevant features. Finally using SVM classifiers, tweets are classified into positive, neutral and negative classes.

3.1. Dataset

The data used in this project is Twitter Dataset. Particularity, SemEval-2014 Task 9 Corpus [**?**] which was collected within the SemEval-2014 Task 9 competition that was a part of the International Workshop on Semantic Evaluation will be used for training and testing. Table 1 shows how the dataset is split into a training set and a testing set.

3.2. Preprocessing Data

A tweet contains a lot of opinions about the data which are expressed in different ways by different users. So keyword extraction is difficult on twitter due to misspellings and slang words. So to avoid this, a preprocessing step is performed before feature extraction. Preprocessing steps include several techniques following to select features from the raw data:

- Remove all URLs (e.g. $www.xyz.com$), hash tags (e.g. $\#topic$), targets (@username), hyperlinks, numbers, repeated letters (more than 2 repeated letters) and punctuation marks.
- Removing stop words that add no sentiment value (articles, some prepositions, etc.).
- Stemming: reducing distinct words to their root form (stem).
- Removing infrequent words (words that appear less than n_{min} number of times).

3.3. Creation of Feature Vector

For the unigram feature, there are usually a larger 500,000 different features. This is a very large number. It makes a model a higher variance. (Since the more complicated model has higher variance). So it will need much more training data to avoid overfitting. Our dataset set contains almost 10 hundreds of sentences. This is a small number of examples. So we need to discard some useless features.

In this project, we use frequency-based feature selection which is the simplest way to do the feature selection. We just pick features (unigram words in our case) for each class with high-frequency occurrence in this class. In practice, if the number of occurrences of a feature is larger than some threshold (100 or 1000 in our experiments), this feature is a good one for that class. As we have seen in the result table, this simple algorithm increases about 0.03 of accuracy.

3.4. Classification Technique

There are different types of classifiers that are generally used for text classification which can be also used for twitter sentiment classification. But in this project, we focus on using Support Vector Machine to do classify twitter sentiment.

SVM Classifier uses a large margin for classification. It separates the tweets using a hyperplane. SVM uses the discriminative function defined as

$$g(X) = w^T \phi(X) + b \quad (1)$$

X is the feature vector, w is the weights vector and b is the bias vector. $\phi()$ is the non-linear mapping from input space to high dimensional feature space. w and b are learned automatically on the training set. Here we used different kernels for classification such as linear and radial basis functions. It maintains a wide gap between two classes

4. Results & discussion

———— Do something here ————————-

5. Conclusions

The SVM has been widely used and promoted for land cover classification studies. So in this project, we proposed the used of SVM with sentiment analysis for classifying the sentence based on twitter data. There are certain issues while dealing with misspellings and slang words from tweets. To deal with these issues, the efficient feature vector is created by doing preprocessing. After preprocessing the raw data, we aim to build an automatic sentiment system by using the twitter dataset which is already labeled. The classification accuracy of the model is tested using the SVM classifier but with a different configuration of parameters. Unigram model is deployed with SVM to classify the sentiment of Twitter data. The training data set can be increased to improve the feature vector related sentence identification process. It may give a better visualization of the content in a better manner that will be helpful for the users. A major conclusion of this study is that the accuracy of SVM classification is influenced by the number of features used.

References

[1] B. Agarwal, V. K. Sharma, and N. Mittal. Sentiment classification of review documents using phrase patterns. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1577–1580. IEEE, 2013.

[2] G. Gautam and D. Yadav. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 437–442. IEEE, 2014.

[3] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009, 2009.

[4] M. Karamibekr and A. A. Ghorbani. Verb oriented sentiment classification. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 327–331. IEEE Computer Society, 2012.

[5] A. Khan and B. Baharudin. Sentiment classification using sentence-level semantic orientation of opinion terms from blogs. In *2011 National Postgraduate Conference*, pages 1–7. IEEE, 2011.

[6] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou. Movie rating and review summarization in mobile environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):397–407, 2011.

[7] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. Evaluation measures for the semeval-2016 task 4: Sentiment analysis in twitter (draft: Version 1.12). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics*, 2016.

[8] L. Ramachandran and E. F. Gehringer. Automated assessment of review quality using latent semantic analysis. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*, pages 136–138. IEEE, 2011.

[9] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.