# Data Representation in Machine Learning-Based Sentiment Analysis of Customer Reviews

Ivan Shamshurin

National Research University – Higher School of Economics, School of Applied Mathematics and Informatics,
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia
ivanshamshurin@gmail.com

**Abstract.** In this paper, we consider the problem of extracting opinions from natural language texts, which is one of the tasks of sentiment analysis. We provide an overview of existing approaches to sentiment analysis including supervised (Naive Bayes, maximum entropy, and SVM) and unsupervised machine learning methods. We apply three supervised learning methods–Naive Bayes, KNN, and a method based on the Jaccard index – to the dataset of Internet user reviews about cars and report the results. When learning a user opinion on a specific feature of a car such as speed or comfort, it turns out that training on full unprocessed reviews decreases the classification accuracy. We experiment with different approaches to preprocessing reviews in order to obtain representations that are relevant for the feature one wants to learn and show the effect of each representation on the accuracy of classification.

**Keywords:** Supervised Learning, Unsupervised Learning, Sentiment Analysis, K-nearest Neighbor, Naive Bayes method, Jaccard index.

## 1 Introduction

People need to collect a lot of information to make the right decision. For instance, when we choose an object $O$ (car, pass to a health resort) we often ask our friends about advantages and disadvantages of $O$. In this case we would value opinions of those people, who used $O$ in the past.

It should be noted that subjective estimations from our friends or $O$-object customers are extremely important because potentially these estimations contain more detailed, comprehensive information about positive and negative features of the goods. It arises from the fact that information from official producers often directs our attention to the advantages of the object $O$.

Generally, sentiment classification in Internet data is an essential modern trend in the text mining [10]. As a result, a program for getting opinion-based estimates of the car was developed.

There are two fundamental principles: supervised learning and unsupervised learning. More detailed information will be given in the next sections of the paper.

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 254–260, 2011.

### 1.1 The Subject of Research and Problems of Sentiment Analysis

In [1] there is an overall survey of the sentiment analysis problems. Themes of tonality defining (a review or comment contains positive, negative or neutral evaluation of a car, film, etc.), objects features detection, summing up opinions, finding out dependent words are analyzed. First of all we are interested in topics related to the tonality defining.

In sentiment analysis a term "object" is used to represent what we attempt to evaluate. In the case of car evaluation the object is a car or its components and properties: saloon, economy of fuel, etc.

Let us consider the problems in sentiment analysis in the following fragment of review:

*"(1) The dynamics is 140 h.p.; it is very good, people respect my car on the road. (2) The ergonomics is 5+. (3) The outside appearance left rivals trailing far behind. (4) The quality of sound insulation is insufficient. (5) In conclusion I have to say that this car is very nice".*

First of all we have to decide what we want to understand from this review. In this text there are sentences with positive valuation (1, 2 and 3) and sentences with negative valuation (4). Also for every opinion there is a corresponding hypotactic object. For instance, in the sentence (5) the author says about the car as a whole, the sentences (1)–(4) are about the dynamics, ergonomics, outside appearance and the sound insulation correspondingly.

The problem of the implicit object description was investigated in [3]. In a car review we can find sentences like this one: "The fuel economy is great!". In this case we see the explicit cars feature as fuel economy valuation. At the same time there are sentences with implicit feature valuation: "All information can be seen". In this case it is more difficult to define that this positive valuation is about the speedometer.

In distinction from [3] in the current paper we took into account implicit valuations in certain experiments. After the analysis of examples of valuations we have an idea about problems, which appear in sentiment analysis. In the current study the following model will be considered: the object is a car; the tonality will be defined for every car feature: "outside appearance", "comfort", "safety", "reliability", "running characteristics".

## 2 Approaches to Sentiment Analysis

In the section 11 of [4] three supervised learning methods were compared on the Internet Movie Database (IMDb): Naive Bayes, Maximal Entropy and SVM methods. The classifiers chose a class (positive or negative) of a review and the accuracy was 81.5%, 81.0% and 82.9% correspondingly.

Contrary to [1], we do not define the tonality of the whole text about a car, but classify it positive or negative or neutral for every aspect of the car. In this sense we are much closer to [3].

In this research authors propose the unsupervised learning for tonality defining for every feature of the object. Contrary to [3] the speciality of the considering task is the fact that we know sentence valuations (i.e. tonality) in advance. This circumstance makes our task more difficult and will decrease the accuracy of the classification, because Internet users can write reviews contradictory to the estimates.

Thus our task is not to define the whole tonality of the car review, but to define the author′s opinion about car features. It is important to note that reviews are not well structured, the style of narration is free. Some features may be not mentioned in the text. If the review is positive in general, it means that the author does not associate the missing feature with negative emotions and we can tag this feature as positive. That is why the task of defining the tonality of the whole text is a subproblem of our task.

## 3 Empirical Protocol

### 3.1 Data Collection and Preprocessing

The Supervised Learning needs the database of the opinions with known cars feature estimates. The comparative analysis of the car-related sites (in Russian) shows that:

- part of them contains opinions without estimates
- there are few sites with opinions with estimates

So, we decided to extract opinion database from the site `http://auto.ru`, because this site is well-known, has high reputation and popularity. Every opinion is placed on the individual page and consists of the title, text, feature′s estimates, advantages and disadvantages.

For automatic opinion extraction the programming module was implemented on the Python. As a result, 5098 opinions about 33 car brands were collected.

Every review was transformed into the following structure: text, then the user′s estimates.

The cross-validation technique [17] was used for obtaining the training and test sets: 5033 reviews were divided $N$ times ($N$ was approximately 100) into the training set by randomly selecting 90% of all the reviews and the test set (10% of all the reviews). The resulting error was considered to be the average of the errors in each partition.

### 3.2 Positive and Negative Examples

We should map scores to classes because the source scores are in the five-point scoring system, and we divide opinions into 2 (positive and negative) and 3 (positive, negative, neutral) classes.

$$\{scores\} \rightarrow \{classes\}$$

**Mapping 1**
Let us define the opinion as positive if it has the scores 4 or 5 and negative if it has the scores 1, 2 or 3.

$$\{1, 2, 3\} \rightarrow \{negative\} \quad \{4, 5\} \rightarrow \{positive\}$$

But as a rule the opinions with score 4 contain pros and cons, so it is preferable to use opinions with score 5 in supervised learning.

**Mapping 2**
The second type of mapping supposes classifying opinions into three classes:

$$\{4, 5\} \rightarrow \{positive\} \quad \{3\} \rightarrow \{neutral\} \quad \{1, 2\} \rightarrow \{negative\}$$

### 3.3 Data Representation

We need to transform the original text into 4 different types of text representation, with increasing level of linguistic processing. It was done with "pymorphy" - the Python library for morphological analysis of texts in Russian.

**Text representation 1:** Text of the opinion without digits, punctuation marks, words consisting of one or two letters, latin symbols.

**Text representation 2:** Text without pronouns, numerals, prepositions, disjunctive and coordinating conjunctions and parenthesises.

**Text representation 3:** Phrases of the following types: noun-adjective (with harmony of tenses, cases, gender).

**Text representation 4:** Normalized adjectives and adverbs.

In all representations the "not" particle is concatenated to the following word.

**Explanations to the text representation 3.** After every phrase we wrote its concatenation, because some of the adjectives are both positive and negative: "a fast car" is a positive characteristic, "fast conked" is a negative one. So, we have 5 attributes for learning: "fast", "car", "conk", "fastcar", "fastconk".

## 4 Descriptions of the Methods

### 4.1 Naive Bayes Method

Naive Bayes Method is a classical machine learning method. We can find its implementation in [2]. Since the probability of a document is a small quantity, these measures were computed on a logarithmic scale.

### 4.2 The Method Based on the Jaccard Measure of Set Similarity

As a result of Naive Bayes learning for every word we have got the conditional probabilities of its appearance in every class. We propose the following method

of Jaccard measure preprocessing: we can set the thresholds $\alpha_1$ and $\alpha_2$, so that three sets of words will be found: $G$, $N$ and $B$.

$$G=\{word \mid P(category = \text{“}good\text{”} \mid word) \geq \alpha_1 P(category = \text{“}neutral\text{”} \mid word)$$

$$\&P(category = \text{“}good\text{”} \mid word) \geq \alpha_2 P(category = \text{“}bad\text{”} \mid word)\}$$

Likewise we find $N$ and $B$. Then for every opinion we define $T$ as the set of words of the opinion. The next stage is the Jaccard measure calculation [4] for pairs $G$ and $T$, $N$ and $T$, $B$ and $T$:

$$M_1 = \frac{|T \cap G|}{|T \cup G|} \quad M_2 = \frac{|T \cap N|}{|T \cup N|} \quad M_3 = \frac{|T \cap B|}{|T \cup B|}$$

If the $M_1 = max\{M_1, M_2, M_3\}$ then we tag the opinion as the opinion with positive estimation of the feature (for example, comfort).

### 4.3 K-Nearest Neighbor

Every document in the analysis can be represented as the frequency vectors of the manually created terms with known tonality:

- positive
fantastic, safety, powerful, etc (83 words in Russian)
- negative
problem, bad, poor, etc (84 words in Russian)

For each opinion we assign the term frequency vector. We took words from the documents with scores 1 (extremely negative) or 5 (positive) with the weight 2.

When the classifier takes an opinion from the test data, we construct the term frequency vector and find $k$ nearest neighbors from the train set. We used the Euclidean metric as the measure of vector similarity.

## 5 First Results

We know the true estimates from the data (each review contains user′s estimate). The machine learning methods provide us with the experimental estimates. So we can compute the precision and recall [8]. It will be the criterium of the method′s accuracy.

On the “positive”-“negative” classification (mapping 1) the most accurate results were achieved with the text representation 3 and 4: 63.3% for Bayes classifier and 56.6% for KNN and 68.1% for Jaccard method. In 3-class classification (mapping 2) the similar precisions were achieved.

The poor precision can be caused by the following noise factors:

- Not always users′ estimates correspond to the text: there are some opinions with scores “1”, but the text of the opinion is extremely positive
- Sometimes in spite of car feature disadvantages in the text, the score is “5”
- Sentences which are not related to the car features are encountered in opinions. People describe the prehistory of the car purchase, for example.

For these reasons the train set should contain only the sentences with the opinions about the cars features. It will decrease the noise influence and more accurate results may be achieved. For more details see section 6.

## 6 Learning with Advanced Data Representation

A list of terms-indicators was made. These 33 words include five cars features and contain other words which are similar to them: "bracket", "dynamics", "automatic gearbox", "body", "clearance", etc.

The most accurate result was achieved on nouns, adjectives, participles, verbs and adverbs from the sentences with the words from the list of terms-indicators.

The results of the sentences with terms-indicators learning, 3-class classification:

| **Bayes** | **KNN ($k$=5)** | **Jaccard** |
|---|---|---|
| Reliability | Running Characteristics | Reliability |
| $Precision = 0.77$ | $Precision = 0.49$ | $Precision = 0.68$ |
| $Recall = 0.77$ | $Recall = 0.36$ | $Recall = 0.44$ |
| Comfort | | Comfort |
| $Precision = 0.66$ | | $Precision = 0.65$ |
| $Recall = 0.66$ | | $Recall = 0.32$ |
| Running Characteristics | | Running Characteristics |
| $Precision = 0.72$ | | $Precision = 0.62$ |
| $Recall = 0.72$ | | $Recall = 0.3$ |

## 7 Conclusion and Future Work

The comparative analysis of the machine learning methods in sentiment analysis was performed. The database of the Internet users′ opinions was extracted. Three methods for cars features estimates defining were tested on it. In general, we see a significant difference in classification accuracy depending on preprocessing reviews.

It stands to mention that the precision values are not very high (65%-70%), because of the noise in data, specific nature of the opinion texts: free style of narration, grammar mistakes.

In the future there is going to be more thorough linguistic processing for decreasing influence of the noise, in particular, tools for spelling correction, words relation defining and removing homonymy are going to be implemented.

## References

1. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg (2006)
2. Segaran, T.: Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly, Sebastopol (2007)

3. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Washington, USA (2004)
4. Berry, M.W., Browne, M.: Lecture notes in data mining. World Scientific Publishing Co, Singapore (2007)
5. Giudici, P.: Applied Data Mining. Statistical Methods for Business and Industry. Wiley, Chichester (2003)
6. Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pp. 174–181. ACL, New Brunswick
7. Lutz, M.: Programming Python. O'Reilly, Sebastopol (2010)
8. van Rijsbergen, C.V.: Information Retrieval, 2nd edn. Butterworth, London; Boston (1979)
9. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly, Sebastopol (2009)
10. Stavrianoui, A., Andritsos, P., Nicoloyannis, N.: Overview and Semantic Issues of Text Mining. SIGMOD Record 36(3), 23–34 (2007)
11. Poirier, D., Bothorel, C., Boulle, M.: Two possible approaches for opinion analysis in film reviews: statistic and linguistic. In: EMOT-2008: LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology (2008)
12. Williams, G.K., Anand, S.S.: Predicting the Polarity Strength of Adjectives Using-WordNet. In: Third International AAAI Conference on Weblogs and Social Media (2009)
13. Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
14. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh (2001)
15. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 417–424 (2002)
16. Huang, A.: Similarity Measures for Text Document Clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC 2008, Christchurch, New Zealand, pp. 49–56 (2008)
17. Geisser, S.: Predictive Inference. Chapman and Hall, New York (1993)