Machine Learning Aspect Category Detection for Sentiment Analysis with Co-Occurrence Data.

Huma Fatima
Student, CSE, MJCET, Hyderabad, Telangana, India.
Syed Shabbeer Ahmed
Professor and Associate Head, CSE,MJCET, Hyderabad, Telangana, India.

ABSTRACT-By means of online consumer reviews as electronic word of mouth facilitate purchase-decision making has become gradually more popular. Web provides a widespread source of customer reviews; one can barely read all reviews to acquire a fair evaluation of a product or service. The information that can be obtained from product and service reviews is not only beneficial to consumers, but also to companies. Knowing what has been posted on the Web can help companies improve their products or services. To effectively handle the large amount of information available in these reviews, a framework for the automated summarization of reviews is desirable. A sub-task that is performed by this framework is to give the general aspect categories addressed in review sentences. For which this paper presents two methods; the first method presented is an unsupervised method that applies clustering on co-occurrence frequency data obtained from a corpus to find these aspect categories. The second method, supervised approach gives the co-occurrence among the words with grammatical connection triples and the aspect categories to know the conditional probability and detect aspect clusters. As a result, it gives more appropriate aspects for the reviews taken from the online websites and made easy for the customers and service providers.

KEYWORDS*:* *Customer reviews, Sentiment analysis, Co-occurrence data, Aspect categories, Electronic Word of Mouth (EWoM), Decision making.*

1. INTRODUCTION

WORD of mouth (WoM) has always been significant on consumer decision-making. The term for this extended form of WoM is electronic WoM (EWoM). One of the most important forms of EWoM communication is product and service reviews posted on the Web by consumers. Retail companies, like Amazon and Yelp allow for easy ways to exchange statements about products, services, and brands. Research has shown these reviews are considered more valuable for consumers than market-generated information and editorial recommendations, used in purchase decision-making. The information that can be acquired from product and service reviews is not only helpful to consumers, but also to companies to improve their products and services. To effectively handle this large amount of reviews a framework is design. An important task of this framework is to automatically summarize the reviews to recognize the topic. These topics are fine-grained into aspect-level sentiment analysis. Supervised and unsupervised machine learning approaches are present in the proposed method. The sentences come from customer reviews and should be classified into one or more aspect categories based on its overall meaning. Let's take an example, given the set of aspect categories (memory, battery, power consumption, slow,heat, and anecdotes/miscellaneous), two annotated sentences are as follows.

"Memory storage problem." → (memory)

"It is very slow and battery is weak." → (battery, slow)

As shown in the above examples, aspect categories do not necessarily occur as explicit terms in sentences. While in the first sentence food, is mentioned explicitly, in the second sentence it is done implicitly. All sentences are assumed to have at least one aspect category present. Because it may not always be clear which category applies to a sentence, due to incomplete domain coverage of the categories and the wide variation of aspects a reviewer can use, a "default" category is used. An example of a sentence where a default category is used, is presented below. Here, the second part of the sentence ("but everything else ... is the pits.") is too general to classify it as one of the other categories (i.e., food, service, price, and ambience).

"The food is outstanding, but everything else about this restaurant is the pits." → (food, anecdotes/miscellaneous).

The unsupervised method uses spreading activation on a graph built from word co-occurrence frequencies in order to detect aspect categories.

2. OBJECTIVE

One can hardly read all the reviews to obtain a fair evaluation of a product or service. A text processing framework is design to detect aspect categories, which is useful for online reviews summarization. Based on the clusters formed sentiments polarity is being decided as negative, positive or neutral. The proposed strategies Supervised and unsupervised approaches help out to know sentiment polarity and aspect category of review data.

3. RESEARCH METHODOLOGY

Since most aspect classifications are left certain in text,1 techniques for distinguishing verifiable fine-grained aspects may be utilized for aspect classes too. All things considered, a few deals with verifiable aspect recognition that motivated this paper are examined underneath.

Table 1. Different research approaches

Approach	Performance
supervised machine learning Kobayashi et al. (2006)	Precision: 67.7% Recall: 50.7%
Li et al. (2010)	Precision: 82.6% Recall: 76.2%
Marcheggiani et al. (2014)	Precision: 86.6% Recall: 78.9%
unsupervised machine *learning* Moghaddam& Ester (2011)	Precision: 74.3% Recall: 86.3%
Sauper&Barzilay (2013)	Precision: 89.1% Recall: 93.4%

4. DATA ANALYSIS

The machine learning approaches Supervised and unsupervised learning is applied to the dataset containing reviews of restaurant reviews.

A. SYSTEM ARCHITECTURE

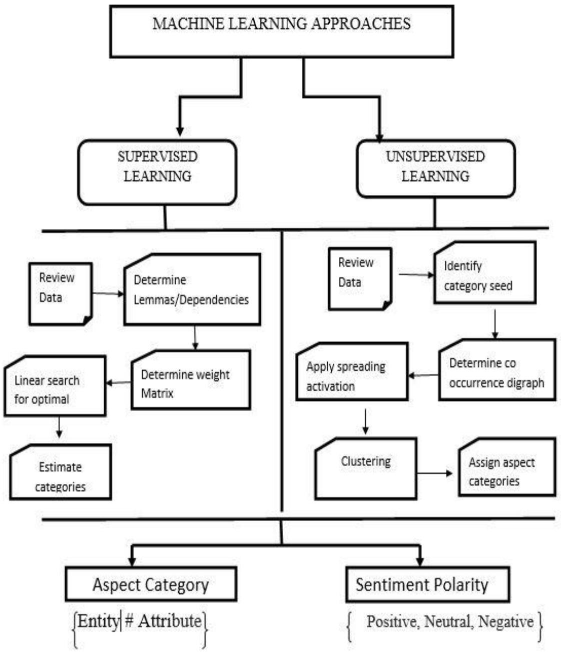


Figure 1: The working scenario of machine learning approaches

B.SUPERVISED LEARNING

The supervised method (called the probabilistic activation method) employs co-occurrence clustering to detect categories.

Similar to lemmas, low frequency dependencies are not taken into account to prevent overfitting, using the parameter αD. Dependencies, describing the grammatical relations between words in a sentence, are more specific than lemmas, as each dependency has three components: 1) governor word; 2) dependent word; and 3) relation type.

The co-occurrence frequencies provide the information needed to find good indicators (i.e., words or dependencies) for the categories. To decide the strength of a pointer, the conditional probability P(B|A) is calculated from the co-occurrence frequency, where category B is disguised when lemma or dependency form A is found in a sentence. These conditional probabilities are easily computed by dividing the co-occurrence frequency of (B,A) by the occurrence frequency of A. The higher this probability, the more likely it is that A implies B. If this value goes beyond a trained threshold, the lemma or dependency form designate the presence of the corresponding category. From fig 1 the supervised working goes in the following way.

1) Determine Lemmas/Dependencies: As a natural language preprocessing step, both training and test data are run through the POS tagger, lemmatizer, and dependency parser of the Stanford CoreNLP.

2)Determine Weight Matrix W: Next all unique categories are identified, storing them in category set C. While the co-occurrence frequencies of all dependency form/lemma-category combinations, are counted and stored in matrix X, respectively.

3) Find Optimal Thresholds: Next we execute a linear search for optimal thresholds. Because the selection of one threshold influences the selection of the other three thresholds, all thresholds are optimized together.

4) Estimate Categories: The final step is to predict the aspect categories for each unseen sentence s∈ test set based on the probability.

Algorithm 1: Identify Category Set C and Compute Weight Matrix W

```
input: training set
input: occurrence threshold θ
output: category set C, Weight matrix W
$C,X,Y \leftarrow \emptyset$
foreachsentence s∈Training set do
// sk are the lemmas/dependecies of s
```

```
foreach sk∈{ sL,sD1,sD2,sD3} do
    foreach dependency forms/lemmas j∈sk do
        // count dependency form/lemma occurrence j in Y
        if j / ∈Y then
            add j to Y
        end
        $Y_j \leftarrow Y_j + 1$
        // sC are the categories of s
        foreach category c∈sC do
            // Add unique categories in category set C
            if c / ∈C then
                add c to C
            end
            // count co-occurrence (c,j) in X
            if (c,j)/ ∈X then
                add (c,j) to X
            end
            $X_{c,j} \leftarrow X_{c,j} + 1$
        end
    end
end
end // Compute conditional probabilities
foreach (c,j) ∈X do
    if $Y_j > \theta$ then
        $W_{c,j} \leftarrow X_{c,j}/Y_j$
    end
end
```

B. UNSUPERVISED LEARNING

The proposed unsupervised method (called the spreading activation method) uses clustering. To avoid having to use the ground truth annotations for this and to keep this method unsupervised, we introduce for each category a set of seed words, consisting of words or terms that describe that category. These words or terms are found by taking the lexicalization of the category, and its synonyms from a semantic lexicon like WordNet.

In our case we want to use spreading activation to find, for each category, a network of words associated with the category's set of seed words. To do this, a network data structure is created, having vertices for all notional words and edges to model the direct relations between these words. In the network data structure all notional words receive an initial activation value of zero except for the category's seed words, which receive positive activation values. From fig 1 the unsupervised working goes in the following way. This algorithm is followed from [1].

1) Identify Category Seed Word Sets: First, we identify for each of the given categories a set of seed words. Containing the category word and any synonyms of that word.

2) Discover Co-Occurrence Digraph: Next, as a characteristic dialect preprocessing step, both preparing and test information are keep running from side to side the lemmatizer of the Stanford CoreNLP. We monitor all lemmas in the content corpus and check their event frequencies. Stop words and lemmas that have an event recurrence lower than a little degree α are disposed of, while whatever is left of the lemmas and relating frequencies are put away in the event vector.

3) Apply Spreading Activation: Once the co-occurrence digraph is obtained, we apply for each category. Each activation value has a range of [0,1], and the closer it is to 1 the stronger the notional word is associated with the considered category.

4) Applying K-means Clustering: K means is an iterative clustering algorithm that aims to find local maxima in each iteration. All notional words are allowed to imply multiple categories except for seed words, which can only imply the category they belong to.

5) Assign Aspect Categories: In the last step we predict categories for each unprocessed sentence, using the clusters K obtained from the previous step. For each unprocessed sentence we use lemmatization, and look if any word matches a rule, after which that cluster is applied.

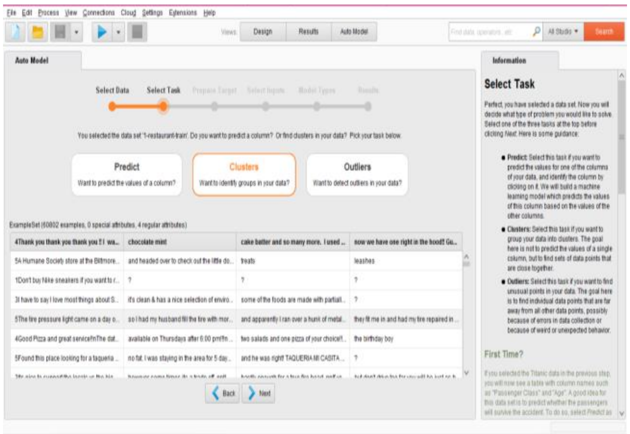
Algorithm2: Spreading Activation Algorithm

```
input : category c
input : vertices V
input : seed vertices Sc
input : weight matrix W
input : decay factor δ
input : firing threshold τc
output: activation values Ac,i for category c
foreach s∈Sc do
    $A_{c,s} \leftarrow 1$
end
foreach $i \in V \setminus Sc$ do
```

```
    | $A_{c,i} \leftarrow 0$
end
F ←Sc
M ←Sc
while M =∅ do
foreach i∈M do
    foreach j∈V do
        | $A_{c,j} \leftarrow \min\{A_{c,j} + A_{c,i} \cdot W_{i,j} \cdot \delta, 1\}$
    end
end
M ←∅
foreach i∈V \F do
    if Ac,i>τ c then
        add i to F
        add i to M
    end
end
end
```

4. EXPERIMENTAL RESULTS

For the evaluation of the proposed methods, the training and test data from SemEval-2014 are used. It contains 3000 training sentences and 800 test sentences taken from restaurant reviews. Each sentence has one or more annotated aspect categories. Fig 2 shows that each sentence has at least one category and that approximately 20% of the sentences have multiple categories. With 20% of the sentences having multiple categories, a method would benefit from being able to predict multiple categories.


File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
Find data, operators...etc
All Studio
Search
Auto Model
Select Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
Information
Back
Next

You selected the data set '1-restaurant-train'. Do you want to predict a column? Or find clusters in your data? Pick your task below.

Predict: Want to predict the values of a column?

Clusters: Want to identify groups in your data?

Outliers: Want to detect outliers in your data?

ExampleSet (60802 examples, 0 special attributes, 4 regular attributes)

4Thank you thank you thank you ! ! wa...	chocolate mint	cake batter and so many more. I used ...	now we have one right in the hood! Gu...
5A Humane Society store at the Biltmore...	and headed over to check out the little do...	treats	leashes
1Don't buy Nike sneakers if you want to r...	?	?	?
3I have to say I love most things about S...	it's clean & has a nice selection of enviro...	some of the foods are made with partial...	?
5The tire pressure light came on a day o...	so I had my husband fill the tire with mor...	and apparently I ran over a hunk of metal...	they fit me in and had my tire repaired in...
4Good Pizza and great service!\nThe dat...	available on Thursdays after 6:00 pm!\n...	two salads and one pizza of your choice!\...	the birthday boy
5Found this place looking for a taqueria ...	no fat. I was staying in the area for 5 day...	and he was right! TAQUERIA MI CASITA ...	?

Select Task

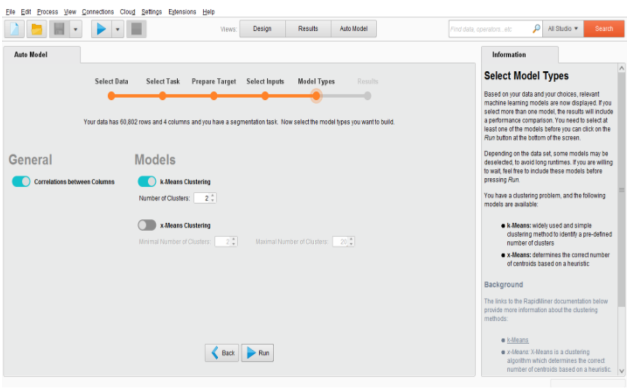
Perfect, you have selected a data set. Now you will decide what type of problem you would like to solve. Select one of the three tasks at the top before clicking Next. Here is some guidance:

- Predict: Select this task if you want to predict the values for one of the columns of your data, and identify the column by clicking on it. We will build a machine learning model which predicts the values of this column based on the values of the other columns.
- Clusters: Select this task if you want to group your data into clusters. The goal here is not to predict the values of a single column, but to find sets of data points that are close together.
- Outliers: Select this task if you want to find unusual points in your data. The goal here is to find individual data points that are far away from all other data points, possibly because of errors in data collection or because of weird or unexpected behavior.

First Time?

If you selected the Titanic data in the previous step, you will now see a table with column names such as "Passenger Class" and "Age". A good idea for this data set is to predict whether the passengers will survive the accident. To do so, select Predict as

Figure 1: Applying clusters on the given data



Your data has 60,802 rows and 4 columns and you have a segmentation task. Now select the model types you want to build.

Correlations between Columns: on

k-Means Clustering: on; Number of Clusters: 2

x-Means Clustering: off; Minimal Number of Clusters: 2; Maximal Number of Clusters: 20

Select Model Types

Based on your data and your choices, relevant machine learning models are now displayed. If you select more than one model, the results will include a performance comparison. You need to select at least one of the models before you can click on the *Run* button at the bottom of the screen.

Depending on the data set, some models may be deselected, to avoid long runtimes. If you are willing to wait, feel free to include these models before pressing *Run*.

You have a clustering problem, and the following models are available:

- k-Means: widely used and simple clustering method to identify a pre-defined number of clusters
- x-Means: determines the correct number of centroids based on a heuristic

Background

The links to the RapidMiner documentation below provide more information about the clustering methods:

- k-Means
- x-Means: X-Means is a clustering algorithm which determines the correct number of centroids based on a heuristic.

Figure 2: Distribution of number of aspect categories persentence.

With 20% of the sentences having multiple categories, a method would benefit from being able to predict multiple categories. This is one of the reasons why clustering is useful in this scenario as multiple clusters can apply to a single sentence.

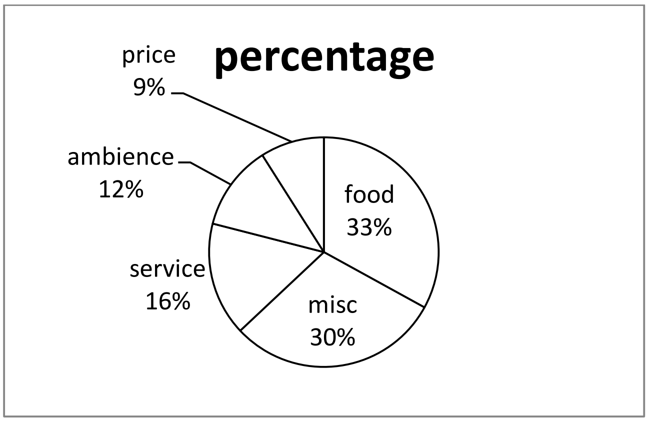


Figure 3: Relative frequency of the aspect categories

Fig. 2 presents the relative frequency of each aspect category, showing that the two largest categories, food and anecdotes/miscellaneous, are found in more than 60% of the sentences. This should make these categories easier to predict than the other categories, not only because of the increased chance these categories appear, but also because there is more information about them.

Table 2: RELATIVE CHANGE IN F1,WHEN VARYING FIRING THRESHOLDS

category	-0.05	-0.02	-0.01	0	0.01	0.02	0.05
Food	-8	-7.9	-7.9	0	-1.9	-6.7	-25
service	-8.6	-3.3	-4.8	0	0	0	0
ambience	-67	-3.1	8.9	0	0	0	-5.6
Price	-72.1	-18.1	-11	0	0	0.1	1.6

Table 2 shows this sensitivity of the firing thresholds, where the relative change in terms of F1-score is given when deviating from the chosen thresholds. As can be seen the proposed method is sensitive to threshold variations.

In Fig. 3, F1-scores are shown for different sizes of the training set, using a stratified sampling technique where the distribution of the categories remains similar to the original data set. Each data point in the figure represents an incremental increase of 10% (300 sentences) in labeled data, for the supervised method, and unlabeled data for the unsupervised method. The supervised method always seems to outperform the unsupervised method, although larger training sizes for the unsupervised method seem to perform on par with the supervised method for which very small amounts of labeled data are available (F1-score around 70%).

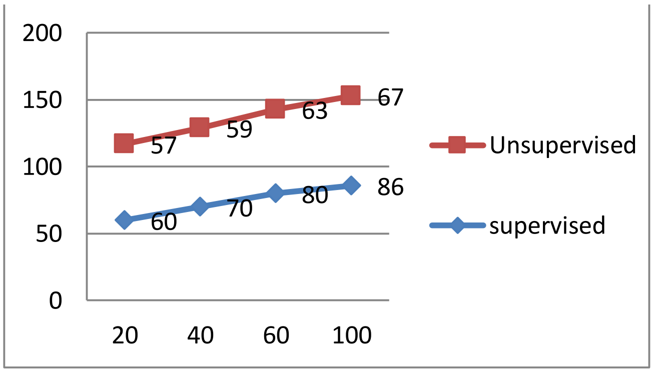


Figure 3: F1-scores for different sizes of the training set (% of 3000 sentence)

5. CONCLUSION

In this paper we have introduced two techniques for identifying perspective classifications that is helpful for online audit synopsis. The first, unsupervised, technique, utilizes spreading initiation over a chart worked from word co-event information, empowering the utilization of both immediate and aberrant relations between words. This outcomes in each word having an actuation esteem for every classification that speaks to the fact that it is so liable to infer that classification. While different methodologies require named preparing information to work, this method works unsupervised. The major drawback of this method is that a few parameters need to be set beforehand, and especially the category firing thresholds (i.e., τc) need to be carefully set to gain a good performance. We have given heuristics on how these parameters can be set. The second, supervised, technique utilizes a somewhat direct co-event strategy where the co-event recurrence between clarified angle classes and the two lemmas and conditions is utilized to figure contingent probabilities.

6. FUTURE WORK

The future workcan be done through forming different clusters on different data to obtain implicit and explicit co-occurrence data. This can give various aspect categories for large datasets. Thus, various clustering can be applied on the data through deep learning.

7. REFERENCES

[1] Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis With Co-Occurrence Data," IEEE Trans. Cybern. Data Eng., vol. 13, pp. 2168-2267 , May 2017

[2] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 813–830, Mar. 2016.

[3] X. Zheng, Z. Lin, X. Wang, K.-J. Lin, and M. Song, "Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification," Knowl. Based Syst., vol. 61, no. 1, pp. 29–47, 2014

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.

[5] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 3, pp. 397–407, May 2012