**Sentiment Analysis in Tourism: Capitalizing on Big Data**

**Background: The Digitally Supported Tourism Industry**

Sentiment analysis basically refers to the use of computational linguistics and NLP to analyze text and identify its subjective information. The interest is driven by (1) escalation of web- and social-media-based information, (2) evolution of new technologies, especially machine learning approaches for text analysis, and (3) development of new business models and applications that make use of this information.

**What Is Sentiment Analysis?**

*Overview*

Sentiment analysis, in particular in relation to customer reviews, is built on the premise that information provided through text (e.g., a review) is either subjective (i.e., opinionated) or objective (i.e., factual). Subjective reviews are

based on opinions, personal feelings, beliefs, and judgment about entities or events. Objective reviews are based on facts, evidences, and measurable observations

Methodologically, sentiment analysis represents a polarity classification problem.

Reviews may not always be subjective, therefore, the binary classification needs to be extended to a ternary classification that contains a third, “objective” category.

Neutral polarity is sometimes interpreted as a polarity between positive and negative.

In sentiment analysis, it is also important to understand what a sentiment relates to. The detection of a target and aspect (i.e., topic detection; Menner et al. 2016), relates to determining the subject of a sentiment expression.

Sentiment analysis can be employed at the word, sentence, paragraph, and document levels.



*Sentiment Analysis Methods*

Sentiment analysis comprises a multistep process: (1) data retrieval, (2) data extraction and selection, (3) data preprocessing, (4) feature extraction, (5) topic detection, and (6) data mining process.

Data retrieval requires the identification and definition of the data source, for example, a commercial service provider portal or a social media network. To collect the review data from these sources, a specific web crawling mechanism is necessary to fetch data and then save them in a database considering the format of data.

Different tasks including splitting a review into sentences, splitting a sentence into words, tokenisation, filtering of stop-words, part-of-speech (POS) tagging, stemming, and the transformation to lower/upper cases are performed on reviews in the preprocessing step to prepare them for the next step.

Feature extraction is known as the process of deriving a set of discriminative, informative and nonredundant values to numerically represent a review or text.

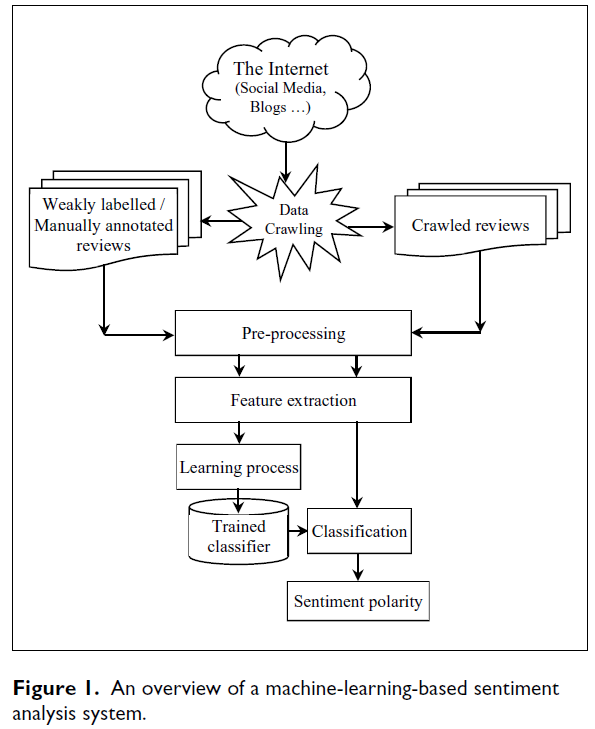
Topic detection is a multiclass classification problem where a text is classified to an appropriate topic class based on its content and application.

In the data mining process, different types of sentiment analysis methods can be distinguished in the literature, namely (1) machine learning, (2) rule-/dictionary-based, and (3) hybrid approaches.

*Supervised Machine Learning Approach*

A sentiment analysis method based on supervised machine learning involves creating a model by using annotated data or weakly labeled corpora.

Supervised machine learning approaches follow several steps (Figure 1). After applying preprocessing techniques to clean, segment and tokenize the text data, a feature extraction method is applied to characterize the review. Features extracted from the reviews are then fed to a classifier to train the classifier. The trained classifier is finally used to determine the polarity of new text.

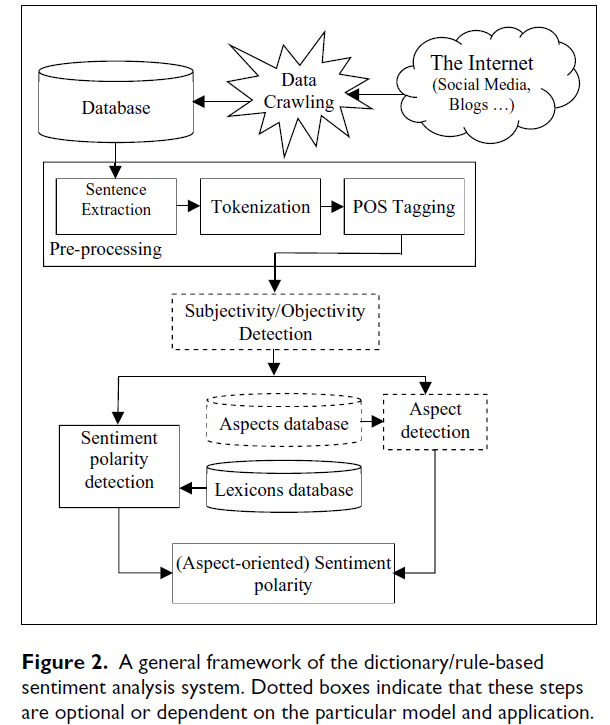


*Unsupervised Machine Learning Approach*

Cluster analysis, as an unsupervised machine learning approach, has been used for data mining, pattern recognition, and image analysis. Clustering is the task of grouping a set of data in such a way that items in a cluster are more similar to each other compared to those in other clusters.

*Dictionary-Based Approach*

Dictionary-based systems rely on the use of comprehensive sentiment lexicons and sets of fine-tuned rules.



*Semantic Approach*

The semantic approach is mainly a rule-based linguistic model to obtain a polarity for each text segment. In this approach a dictionary of domainspecific terms and their associated polarity values is required.

*Hybrid Approach*

In hybrid approaches, dictionary and machine learning-based approaches can work in parallel to compute two sentiment polarities.

**Review of Tourism-Focused Sentiment Analysis**

*Identified Studies and Datasets Used*

*Evaluation Metrics*

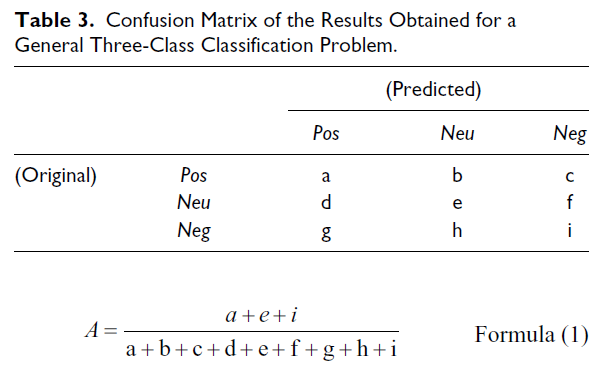
As mentioned earlier, most sentiment analysis methods provide either a two-class (positive and negative) or a threeclass (positive, neutral, and negative) classification. A clean and unambiguous way to present the prediction results of a classifier is to use a confusion matrix, which is also called contingency table

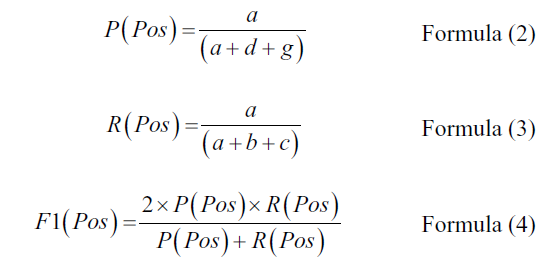
The Accuracy (*A*) is one of the evaluation metrics commonly used in the literature. It is simply the number of correct predictions of sentiment made, divided by the total number of predictions made. The accuracy measures how accurate the method is in its prediction of the correct output.

Precision, Recall, and *F*1-measure are the other three evaluation metrics frequently used for evaluating the results of sentiment analyses.

Considering a sample sentiment analysis system of three classes, and using the definitions provided in Table 3, the Precision (*P*) of a class, for example positive (*Pos*), is defined as the ratio of the number of instances correctly classified as the class *Pos* relative to the total number of instances predicted as the class *Pos*. The Recall (*R*) of a class, for example *Pos*, is then defined as the ratio of the number of instances correctly classified as the class *Pos* with respect to the total number of instances, which actually should be classified as the class *Pos*. The *F*1 measure is a weighted harmonic mean of both, the Precision and Recall.

The described metrics for the three-class problem can easily be adapted for the two-class problem by removing the Neutral column and row from Table 3. Based on the abovementioned definitions, the *P, R*, and *F*1 measures of the *Pos* class are computed as follows:





*Performance of Sentiment Analyses in Tourism Studies*

*Synthesis of the Tourism-Specific Results*

From the results reported in the literature, we noted that most sentiment analysis methods perform better in classifying positive sentences than negative or neutral sentences.

**Concluding Remarks and Future Directions**

Due to the difficulty of detecting and finding implicit aspects in reviews, aspect-oriented sentiment analysis is still a challenging problem.

**Declaration of Conflicting Interests**