Survey on Aspect-Level Sentiment Analysis

Kim Schouten and Flavius Frasincar

Abstract—The field of sentiment analysis, in which sentiment is gathered, analyzed, and aggregated from text, has seen a lot of attention in the last few years. The corresponding growth of the field has resulted in the emergence of various subareas, each addressing a different level of analysis or research question. This survey focuses on aspect-level sentiment analysis, where the goal is to find and aggregate sentiment on entities mentioned within documents or aspects of them. An in-depth overview of the current state-of-the-art is given, showing the tremendous progress that has already been made in finding both the target, which can be an entity as such, or some aspect of it, and the corresponding sentiment. Aspect-level sentiment analysis yields very fine-grained sentiment information which can be useful for applications in various domains. Current solutions are categorized based on whether they provide a method for aspect detection, sentiment analysis, or both. Furthermore, a breakdown based on the type of algorithm used is provided. For each discussed study, the reported performance is included. To facilitate the quantitative evaluation of the various proposed methods, a call is made for the standardization of the evaluation methodology that includes the use of shared data sets. Semantically-rich concept-centric aspect-level sentiment analysis is discussed and identified as one of the most promising future research direction.

Index Terms—Text mining, Linguistic processing, Machine learning, Text analysis, Sentiment analysis, Aspects

1 Introduction

The digital age, also referred to as the information society, is characterized by ever growing volumes of information. Driven by the current generation of Web applications, the nearly limitless connectivity, and an insatiable desire for sharing information, in particular among younger generations, the volume of user-generated social media content is growing rapidly and likely to increase even more in the near future. People using the Web are constantly invited to share their opinions and preferences with the rest of the world, which has led to an explosion of opinionated blogs, reviews of products and services, and comments on virtually anything. This type of web-based content is more and more recognized as a source of data that has added value for multiple application domains.

1.1 Applications

For ages, governments and mercantile organizations alike have been struggling to determine the opinions of their target communities and audiences. Now, for the first time, people voluntarily publish their opinions on the World Wide Web, for anyone to see. This social Web allows for almost immediate feedback on products, stocks, policies, etc., and many of the desired data, which was hard to come by in the past, is now readily available. This is in stark contrast with the traditional surveys and questionnaires that often reluctant participants had to fill without any personal motivation to do so, resulting in sub-optimal information.

Many individuals are influenced by the opinionated materials they find on the Web. This is especially true for product reviews, which have been shown to influence buying behavior [1]. Moreover, information provided by individuals on the Web is regarded as more trustworthy than information provided by the vendor [1]. From a producers point of view, every person is a potential customer. Hence, knowing their likes and dislikes can be of great help in developing new products [2], as well as managing and improving existing ones [3]. Furthermore, understanding how the information in, for example, product reviews interacts with the information provided by companies enables the latter to take advantage of these reviews and improve sales [4]. In fact, opinions on the Web have become a resource to be harnessed by companies, just like the traditional word-of-mouth [5]. In addition to this traditional producer/consumer model, sentiment analysis is also important for other economic areas, like for example financial markets [6].

1.2 Definitions

This survey will start with a quick summary of the definitions for aspect-level sentiment analysis set forth by Pang and Lee [3]. The field of sentiment analysis operates at the intersection of information retrieval, natural language processing, and artificial intelligence. This has led to the use of different terms for similar concepts. A term often used is 'opinion

• *K. Schouten and F. Frasincar are with the Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands*
E-mail: {schouten, frasincar}@ese.eur.nl

mining', a denotation coming from the data mining and information retrieval community. The main goal of opinion mining is to determine the opinions of a group of people regarding some topic. The term 'sentiment analysis' is also used quite often. It comes from the natural language processing domain, and the focus lies on determining the sentiment expressed in text. The term subjectivity analysis is sometimes regarded as encompassing opinion mining and sentiment analysis, as well as related tasks [7], but also as a sub-task of opinion mining and sentiment analysis [8]. Nevertheless, all these terms, even though possibly used for slightly different tasks or different angles, represent the same area of research. This field of research, labeled as opinion mining, sentiment analysis, or subjectivity analysis, studies the phenomena of opinion, sentiment, evaluation, appraisal, attitude, and emotion [8]. For ease of reference these terms are often simply referred to as opinion or sentiment, even though they are technically not the same.

An opinion can be defined as a "judgment or belief not founded on certainty or proof" [9]. In this sense, it is the opposite of a fact. Hence, statements expressing an opinion are subjective, while factual statements are objective. Sentiment is orthogonal to this [10], as it is closely related to attitude and emotion, used to convey an evaluation of the topic under discussion. Because of this orthogonality, there are four quadrants a sentence can fall in. It can be subjective or objective, as well as with or without sentiment. For example, people may have varying opinions on what color a certain dress is[1] in "Others think it looks like a blue and black dress, but to me it is a white and gold dress.", without expressing any sentiment. In contrast, the statement "Some persons looked at the dress and saw a blue and black one, others were convinced it was white with gold instead" is purely objective and also without sentiment. Statements conveying sentiment can be both subjective and objective as well. For example "The blue and black dress is the most beautiful" is a subjective statement with sentiment, while "My favorite dress is sold out" is an objective statement with sentiment. In light of the above discussion, we will use the term sentiment analysis throughout this survey, as it best captures the research area under investigation.

With the above discussion in mind, finding sentiment can be formally defined as finding the quadruple (s, g, h, t) [8], where s represents the sentiment, g represents the target object for which the sentiment is expressed, h represents the holder (i.e., the one expressing the sentiment), and t represents the time at which the sentiment was expressed. Note that most approaches focus only on finding the pair (s, g). The target can be an entity, such as the overall topic of the review, or an aspect of an entity, which can be any characteristic or property of that entity. This decision is made based on the application domain at hand. For example, in product reviews, the product itself is usually the entity, while all things related to that product (e.g., price, quality, etc.) are aspects of that product. Aspect-level sentiment analysis is concerned, not just with finding the overall sentiment associated with an entity, but also with finding the sentiment for the aspects of that entity that are discussed. Some approaches use a fixed, predetermined list of aspects, while others freely discover aspects from the text.

Both sentiment and target can be expressed explicitly or remain implicit, independent of each other. When explicitly mentioned, a sentiment or target is literally in the text, while implicit expressions of sentiment or target have to be inferred from the text, which sometimes even requires additional context or domain knowledge. For example, "This hotel is fantastic" is an example of a sentence with an explicit entity and an explicit sentiment, while "The service is great" expresses a similar explicit sentiment, but with an explicit aspect of an entity as its target. On the other hand, "I could not sleep because of the noise" is an example that illustrates an implicit sentiment with an implicit target: one expects to be able to sleep well, but according to the sentence, this expectation was not met, which is why this sentence can be seen as illustrating a negative sentiment.

Last, since the set of human emotions is very large [11], sentiment polarity is often used instead. Polarity describes the direction of the sentiment and it is either positive, negative, or neutral [3]. Some algorithms only perform a binary classification, distinguishing solely between positive and negative polarity.

1.3 Outline of Aspect-Level Sentiment Analysis

In general, three processing steps can be distinguished when performing aspect-level sentiment analysis: identification, classification, and aggregation [7]. While in practice, not every method implements all three steps or in this exact order, they represent major issues for aspect-level sentiment analysis. The first step is concerned with the identification of sentiment-target pairs in the text. The next step is the classification of the sentiment-target pairs. The expressed sentiment is classified according to a predefined set of sentiment values, for instance positive and negative. Sometimes the target is classified according to a predefined set of aspects as well. At the end, the sentiment values are aggregated for each aspect to provide a concise overview. The actual presentation depends on the specific needs and requirements of the application.

Besides these core elements of aspect-level sentiment analysis, there are additional concerns: robustness, flexibility, and speed. Robustness is needed in

1. See for instance http://www.wired.com/2015/02/science-one-agrees-color-dress/

order to cope with the informal writing style found in most user-generated content. People often make lots of errors in spelling and grammar, not to mention the slang language, emoticons, and other constructions that are used to voice a certain sentiment. Flexibility is the ability to deal with multiple domains. An application may be performing very well on a certain domain, but very poorly on another, or just mediocre on all domains. Last, an aspect-level sentiment analysis solution ideally is accessible using a Web interface due to Web ubiquity, underlining the need for high speed performance.

1.4 Focus of this Survey

To allow for a proper level of depth, we focus this survey on a particular sub-field of sentiment analysis. As discussed in [8], sentiment analysis has been studied mainly at three levels of classification. Sentiment is classified on either the document level, the sentence level, or the entity or aspect level. A focus on the first level assumes that the whole document expresses sentiment about only one topic. Obviously, this is not the case in many situations. A focus on the second level comes with a similar assumption in that one sentence should only contain sentiment about one topic. Within the same sentence, it is often the case that multiple entities are compared or that certain sentiment carrying opinions are contrasted. At both the document level and the sentence level, the computed sentiment values are not directly associated with the topics (i.e., entities or aspects of entities) discussed in the text. In a similar manner, sentiment can be computed over any arbitrary piece of text, even a complete corpus (e.g., a corpus of microblog entries, where each post is considered a document).

In contrast, aspect-level sentiment analysis aims to find sentiment-target pairs in a given text (i.e., this could range from sentences or smaller textual units, to complete corpora containing many documents). Within aspect-level sentiment analysis, the overall sentiment would generally refer to the entity, while aspect-level sentiment analysis would refer to the sentiment associated with aspects of the entity being discussed. This allows for a more detailed analysis that utilizes more of the information provided by the textual review. Therefore, this survey will focus on aspect-level analysis and its various sub-tasks. This also allows us to cover more recent developments, instead of repeating established insights that can be found in other surveys [3] [7] [8] [12].

A good survey and introduction into the field of sentiment analysis is Pang and Lee's publication from 2008 [3]. Not only are various techniques and applications discussed, but also ethical, practical, and theoretical considerations are covered by their article. However, the coverage of the survey is restricted mostly to document-level machine learning approaches. There is a smaller survey by Tang et al. [12] from 2009, and while it mainly focuses on document-level machine learning approaches as well, it specifically addresses the domain of consumer reviews. Tsytsarau and Palpanas [7] published a survey in 2011 that, while still focusing on document-level sentiment analysis, distinguishes between four different approaches for identifying the sentiment value of words: machine learning, dictionary-based, statistical, and semantic. These four labels mainly describe how the sentiment value of a single word is determined. In 2012, Liu published a survey [8], with an updated overview of the entire field of sentiment analysis. The chapter dealing with aspect-level sentiment analysis is organized as a list of sub-problems that one encounters when implementing an actual solution: from definitions to aspect extraction, including various challenges that can be defined as part of aspect-level sentiment analysis, like dealing with implicit and explicit sentiment and entities, to how aspects and sentiment values can be identified and linked to one another. However, a systematic classification of approaches and reports of their accuracy are missing, a gap that the current survey is aiming to fill.

1.5 Organization

This survey is organized as follows. First, we discuss the evaluation methodology for aspect-level sentiment analysis. Then, we present various approaches for aspect detection and sentiment analysis in isolation as well as joint aspect detection and sentiment analysis approaches. After that, we discuss some interesting related problems that most approaches encounter and present some solutions dedicated to solve these issues. Then, the problem of aggregating sentiment scores is discussed, as well as the presentation of the aspect sentiment scores. We conclude the paper with an informed outlook on the field of aspect-level sentiment analysis and highlight some of the most promising directions for future research.

2 Evaluation Methodology

Any maturing research area has to arrive at a common evaluation methodology that is generally accepted in the field. For aspect-level sentiment analysis, this is not yet the case, as evidenced by the wide variety of used evaluation measures and data sets.

In recent years, the International Workshop on Semantic Evaluation has embraced the task of aspect-level sentiment analysis [13] [14], providing a controlled evaluation methodology and shared data sets for all participants. All competing systems get the same unannotated test data, which they will have to annotate with aspect tags and sentiment tags. This is sent to the organization which will perform a controlled evaluation of the provided data using the same procedures for each competing system. The result

is an overview of approaches that can be directly compared against each other.

Likewise, the GERBIL framework [15] also has the goal of directly comparing approaches with the same, controlled, evaluation methodology. To that end, it combines multiple data sets and many implementations of existing algorithms to compare against. Furthermore, the exact experimental setting is permanently stored and can be referred to so that readers can exactly see how the evaluation is performed. Unfortunately, at the time of writing, this system is only available for the task of entity annotation. However, the concept is applicable to many tasks, including aspect-level sentiment analysis.

Of course, many problems arise when research field standards are developed. For instance, the annotations needed differ for the various approaches since some methods classify sentiment in only positive or negative, while others use a five-star rating. In other cases, the specific focus of an evaluation may not be aspect-level sentiment analysis, like in [16] where the task of selecting comprehensive reviews is evaluated. The focus on different tasks also solicits the use of a wide variety of evaluation metrics.

2.1 Evaluation Measures

Currently, most of the surveyed work uses accuracy, precision, recall, and F_1 to measure quantitative performance, but some less common metrics are in use as well. To facilitate the proper interpretation of the reported performances, we will briefly discuss these less common metrics and present the general way of computing them.

For sentiment classification, multiple measures are in use: Ranking Loss, Mean Absolute Error, and Mean Squared Error. All of them assume that the sentiment value is at least an interval type variable. This assumption can be reasonable, even though in practice this is usually not the case.

Ranking Loss [17], used in [18], measures the average distance between the true rank and the predicted rank. For a sentiment classification problem with m sentiment classes (e.g., on a scale from one to five) and n test instances, Ranking Loss is defined in Equation 1 as the average deviation between the actual sentiment value y for instance i and the predicted sentiment value $\hat{y}$ for that instance.

$$\text{Ranking Loss} = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{m \times n} \tag{1}$$

An alternative to Ranking Loss is the macro-averaged Mean Absolute Error, which is particularly robust to imbalance in data sets. Used in [19], it is computed as

$$\text{MAE}^M(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{|\mathbf{y}_j|} \sum_{y_i \in \mathbf{y}_j} |y_i - \hat{y}_i| \tag{2}$$

where $\mathbf{y}$ is the vector of true sentiment values, $\hat{\mathbf{y}}$ is the vector of predicted sentiment values, $\mathbf{y}_j = \{y_i : y_i \in \mathbf{y}, y_i = j\}$, and m is the number of unique sentiment classes in $\mathbf{y}$.

A similar measure is Least Absolute Errors (LAE), or L_1 error, which is used in [20] to measure sentiment classification error. It is computed as

$$\text{LAE} = \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{3}$$

where $\hat{\mathbf{y}}$ is the vector of n sentiment predictions and $\mathbf{y}$ is the vector of true sentiment values.

Related to this is the Mean Squared Error (MSE), or the mean L_2 error, used in [21] to evaluate the sentiment prediction error of the proposed method. This is a widely used metric, especially for regression, which is computed as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{4}$$

where, again, $\hat{\mathbf{y}}$ is the vector of n sentiment predictions and $\mathbf{y}$ is the vector of true sentiment values.

For aspect detection, some algorithms return a ranked list of aspects. To compare rankings, multiple measures exist, one of which, the normalized Discounted Cumulative Gain, is used when reporting performance scores for the discussed work.

The normalized Discounted Cumulative Gain (nDCG) [22], also used in [21], is particularly useful to evaluate relevance for lists of returned aspects. Furthermore, relevance does not have to be binary. The regular Discounted Cumulative Gain is computed as

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^{\text{rel}(i)} - 1}{\log_2(i+1)} \tag{5}$$

where k represents the top k returned aspects that will be evaluated, and $rel(i)$ is the relevance score of aspect i. To normalize this score, and allow cross-query evaluation, the DCG score is divided by the ideal DCG. This is the DCG that would have been returned by a perfect algorithm. For most of the discussed approaches, nDCG cannot be computed, since it does not return a ranked list. However, if an algorithm produces rankings of aspects, for instance, based on how much these are discussed in a review, nDCG is an effective way of summarizing the quality of these rankings.

When dealing with generative probabilistic models, like topic models, where the full joint probability distribution can be generated, it is also possible to use the KullbackLeibler divergence [23], or KL-divergence for short. This measures the difference between two probability distributions, where one distribution is the one generated by the model and the other is the distribution that represents the true data. How the KL-divergence is computed depends on the exact

situation, for example whether the probability distributions are continuous or discrete. Characteristic for the KL-divergence, compared to other measures is that it is not a true metric, since it is not symmetrical: the KL-divergence of A compared to B is different than the KL-divergence of B compared to A.

3 Core Solutions

To provide insight into the large number of proposed methods for aspect-level sentiment analysis, a task-based top-level categorization is made, dividing all approaches into the following three categories: methods focusing on aspect detection, methods focusing on sentiment analysis, methods for joint aspect detection and sentiment analysis. Within each task, a method-based categorization is made that is appropriate for that task (e.g., supervised machine learning, frequency-based, etc.). For each task, a table outlining all surveyed methods that cover that task is given. Each table lists the work describing the method, its domain (i.e., what kind of data it is evaluated on), a short description of the task that is evaluated, and the performance as reported by the authors. For the methods that perform sentiment analysis, the number of sentiment classes is also reported. Note that since evaluation scores are taken from the original papers, experimental settings will be different for each work and as a consequence the methods cannot be compared using these evaluation scores. When multiple variants of an approach are evaluated and compared, we report only the results of the variant that yields the best performance. When the same method is evaluated over multiple data sets, the results are presented as the average or as a range.

Note that work describing both a method for aspect detection and a different method for sentiment analysis appears twice: the aspect detection method is discussed in Section 3.1, while the sentiment analysis method is discussed in Section 3.2. A tree overview of the classification system is shown in Figure 1, which is inspired by the organization of approaches that is used in the tutorial of Moghaddam & Ester [24].

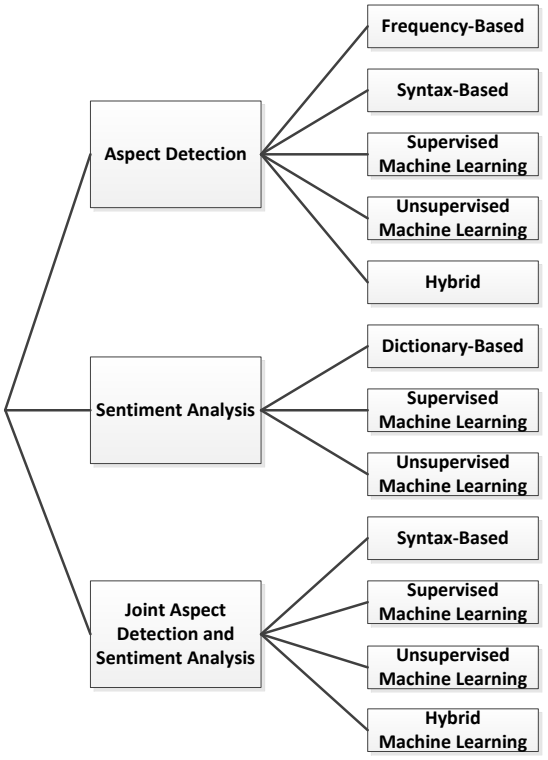

Fig. 1. Taxonomy for aspect-level sentiment analysis approaches using the main characteristic of the proposed algorithm.

3.1 Aspect Detection

All methods featuring an aspect detection method of interest are discussed in this section. A division is made between frequency-based, syntax-based (sometimes referred to as relation-based methods), supervised machine learning, unsupervised machine learning, and hybrid approaches. All the discussed approaches, together with their reported performance can be found in Table 1.

3.1.1 Frequency-Based Methods

It has been observed that in reviews, a limited set of words is used much more often than the rest of the vocabulary. These frequent words (usually only single nouns and compound nouns are considered) are likely to be aspects. This straightforward method turns out to be quite powerful, a fact demonstrated by the significant number of approaches using this method for aspect detection. Clear shortcomings are the fact that not all frequent nouns are actually referring to aspects. Some nouns in consumer reviews, such as 'dollar' or 'bucks', are just frequently used. On the other hand, aspects that are not frequently mentioned, like very specific aspects that most people do not discuss, will be missed by frequency-based methods. To offset these problems, frequency-based methods can be supplemented with a set of rules to account for some of these issues. However, these manually crafted rules often come with parameters which have to be tuned.

The most well-known approach featuring a frequency-based method for aspect detection is [25]. The same authors describe the matching sentiment analysis method in [26], which will be explained in Section 3.2.1. The aspect detection method described in [25] only considers single nouns and compound nouns as possible aspects. First, the frequency of each combination of nouns is retrieved. For this, the nouns do not have to be next to each other, they should just appear in the same sentence. This helps to find aspects like 'screen size' when it is phrased as 'size of the screen'. The noun combinations that occur in at least 1% of the sentences are considered

as aspects. Two rules are used to prune the result in order to lower the number of false positives. The first aims to remove combinations where the nouns never appear closely together, while the second aims to remove single-word aspects which appear only as part of a multi-word aspect. When a sentence does not contain a frequent aspect but does contain one or more sentiment words, as indicated by the sentiment analysis method used in conjunction with the current approach, then the noun or compound noun nearest to the sentiment word is extracted as an infrequent aspect. This process, while sensitive to generating false positives, is able to increase the recall of the method. An improvement to this process can be found in [16], where grammatical dependencies are employed to find infrequent aspects instead of word distance. In this particular study, the goal is to find reviews that are most comprehensive with respect to a certain aspect, so that sentiment analysis can be performed on reviews that have a thorough discussion on that aspect.

Only explicit aspects are detected in [25], but [27] employs association rule mining to find implicit aspects as well. By restricting sentiment words to appear as rule antecedents only, and aspect words to appear as rule consequents, the generated association rules can now be used to find aspects based on already found sentiment words. Last, a major difference between the two methods is that, while [25] generates the frequent item sets from a transaction file, [27] generates its rules from the co-occurrence matrix of the bipartite of sentiment words and explicit aspects. One should note that these explicit features must therefore first be detected, before implicit features can be found using this method. The two methods can thus be thought of as complementary.

Similar to [25] is [28], where a supervised form of association rule mining is used to detect aspects. Instead of the full review text, [28] targets pros and cons that are separately specified on some Web sites. Since pros and cons are known to be rich in aspect descriptions, this task is allegedly simpler than detecting aspects in the full text, and the obtained results are obviously better than those of [25].

A major shortcoming of most frequency-based methods is the fact that nouns and noun phrases that naturally have a high frequency are mistakenly seen as aspects. Red Opal, a system introduced in [29], aims to address this issue by comparing the frequency of a prospective aspect with baseline statistics gathered from a corpus of 100 million words of spoken and written conversational English. To be considered as an aspect, a word or bigram has to appear more often in a review than is likely given its baseline frequency. This improves feature extraction and reduces the number of non-features because these non-features are usually often occurring words that would be above a fixed threshold but are filtered out when using baseline statistics. As part of the evaluation, a small scale survey was conducted to assess the actual helpfulness of the extracted features, which suggested that users prefer bigram features over unigram features and specific features over more generic features. The same concept of baseline statistics is used in [30], where it used to filter the list of high-frequency noun phrases. Additionally, a part-of-speech pattern filter is also applied, such that every aspect needs to be followed by an adjective (note that this filter is designed to work with Chinese texts).

3.1.2 Syntax-Based Methods

Instead of focusing on frequencies to find aspects, syntax-based methods find aspects by means of the syntactical relations they are in. A very simple relation is the adjectival modifier relation between a sentiment word and an aspect, as in 'fantastic food', where 'fantastic' is an adjective modifying the aspect 'food'. A strong point of syntax-based methods is that low-frequency aspects can be found. However, to get good coverage, many syntactical relations need to be described.

To mitigate the low recall problem, a generalization step for syntactic patterns using a tree kernel function is proposed in [31]. Given a labeled data set, the syntactic patterns of all the annotated aspects are extracted. Then, for the unseen data, syntax trees of all sentences are obtained. Instead of directly trying to find an exact match between the aspect pattern and the syntax tree, both are split into several different substructures. Then the similarity between the pattern and a sentence can be measured as the number of matching substructures. The common convolution tree kernel is used to compute similarity scores for each pair of substructures, with a threshold determining whether a pair is a match or not.

In [33] (an extended version was published later in [53]), and its extension [34], aspect detection and sentiment lexicon expansion are seen as interrelated problems for which a double propagation algorithm is proposed, featuring parallel sentiment word expansion and aspect detection. With each extra known sentiment word, extra aspects can be found, and with additional known aspect words, more sentiment words can be found, etc. The algorithm continues this process until no more extra sentiment words or targets can be found. To find sentiment words based on known aspect words, and the other way around, a set of rules based on grammatical relations from the employed dependency parser, is constructed. In this way, more sentiment-aspect combinations can be found and classified in a given text than with previous approaches. A big advantage of this method is that it only needs a small seed set to work properly compared to the large corpus most trained classifiers require.

TABLE 1
Approaches for aspect detection

	domain	evaluation task	performance
frequency-based			
Hu & Liu (2004) [25]	product reviews	aspect detection	precision: 72% recall: 80%
Long et al. (2010) [16]	hotel reviews	comprehensive review selection	F_1: 70.6% - 93.3%
Hai et al. (2011) [27]	cell phone reviews	implicit aspect detection	precision: 76.29% recall: 72.71%
Liu et al. (2005) [28]	product reviews	aspect detection (pros/cons)	precision: 88.9% / 79.1% recall: 90.2% / 82.4%
Scaffidi et al. (2007) [29]	product reviews	aspect detection	precision: 85%-90% complexity: O(n)
Li et al. (2009) [30]	product reviews	aspect detection	F_1: 74.07%
syntax-based			
Zhao et al. (2010) [31]	car, camera, and phone reviews [32]	aspect detection	precision: 73%, 66%, and 76% recall: 63%, 67%, and 68%
Qiu et al. (2009) [33]	product reviews (data from [26])	aspect detection	precision: 88% recall: 83%
Zhang et al (2010) [34]	cars & mattress reviews phone & LCD forum posts	aspect detection	precision: 78% / 77% recall: 56% / 64% precision: 68% / 66% recall: 44% / 55%
supervised machine learning			
Jakob & Gurevych (2010) [35]	data from [36] [37] [38]	opinion target extraction	precision: 61.4% - 74.9% recall: 41.4% - 66.1%
unsupervised machine learning			
Titov & McDonald (2008) [39]	product, hotel, and restaurant reviews	aspect detection	no quantitative evaluation
Lu et al. (2011) [20]	hotel [40] & restaurant [41] reviews	sentence labeling	accuracy: 79.4% F_1: 71.4% - 85.6%
Lakkaraju et al. (2011) [42]	product reviews	aspect detection (2/5-class)	precision: 83.33% / 82.52% recall: 81.12% / 80.72%
Zhan & Li (2011) [43]	hotel [44] & restaurant [41] reviews		not available (graphs only)
Wang et al. (2011) [21]	hotel [40] & mp3 player reviews	aspect rating prediction	MSE: 1.234 nDCG: 0.901
Moghaddam & Ester (2013) [45]	product reviews [46] [40] [47]	item categorization (cold) item categorization (default)	accuracy: 79%-86% accuracy: 95%-97%
Hai et al. (2014) [48]	product reviews	aspect detection	no quantitative analysis
hybrid			
Popescu & Etzioni (2005) [49]	product reviews (data from [26])	aspect detection	precision: 87.84% recall: 77.6%
Yu et al. (2011) [50]	product reviews	aspect detection	F_1: 70.6% - 76.0%
Raju et al. (2009) [51]	product reviews	aspect detection (incl. partial matches)	precision: 92.4% recall: 62.7%
Blair-Goldensohn et al. (2008) [52]	restaurant & hotel reviews	static aspect detection	precision: 70.5% - 94.6% recall: 47.1% - 82.2%

3.1.3 Supervised Machine Learning Methods

There are not many supervised machine learning methods for aspect detection that are purely machine learning methods. Since the power of supervised approaches lies in the features that are used, feature construction often consists of other methods (e.g., frequency-based methods) in order to generate more salient features that generalize better than simple bag-of-words or part-of-speech features.

In [35], aspect detection is cast as a labeling problem, which is solved by using a linear chain Conditional Random Field (CRF), common in natural language processing, to process a whole sequence (e.g., a sentence) of words. This automatically takes the context of a word into account when assigning it a label. Multiple features are used when determining the best label for a word, including the actual word, its part-of-speech tag, whether a direct dependency relation exists between this word and a sentiment expression, whether this word is in the noun phrase that is closest to a sentiment expression, and whether this word is in a sentence that actually has a sentiment expression. The ground-truth from a subset of the used data sets [36] [37] [38] is used to train the model. Four domains are covered in these review data sets: movies, web-services, cars, and cameras.

3.1.4 Unsupervised Machine Learning

In general, this class of models operates unsupervised, requiring only labeled data to test and validate the model. Nevertheless, a large amount of data is generally needed to successfully train these type of models. Most of the approaches in this section use LDA, which is a topic model proposed in [54]. Each document

is viewed as a mixture of topics that could have generated that document. It is similar to probabilistic Latent Semantic Analysis [55] but it utilizes a Dirichlet prior for the topic distribution instead of a uniform topic distribution. One of the main drawbacks of LDA is that the generated topics are unlabeled, preventing a direct correspondence between topics and specific aspects or entities. And while sometimes a quick glance at the words associated with a topic is enough to deduce which aspect it is referring to, not all topics are that clear cut. Because LDA utilizes a bag of words approach when modeling documents and topics, the contents of a topic (i.e., the words associated with it) are not required to be semantically related: it might be impossible to characterize a topic, making it much less suitable for interpretation.

Since LDA was designed to operate on the document level, employing it for the much finer-grained aspect-level sentiment analysis is not straightforward. Some critical issues that arise when implementing an LDA-based method for aspect-level sentiment analysis have been discussed in [39]. The main argument is that since LDA uses a bag of words approach on the document level, it will discover topics on the document level as well. This is good when the goal is to find the document topic (i.e., this could be the entity, or some category), but not as useful when one is looking for aspects. The topics that LDA returns are simply too global in scope to catch the more locally defined aspects. One way to counter this would be to apply LDA on the sentence level, but the authors argue that this would be problematic since the bag of words would be too small, leading to improper behavior of the LDA model (cf. [56]). Although some solutions exist to this problem in the form of topic transitions [57], the authors deem those computationally too expensive. Instead an extension to LDA is proposed called Multi-grain LDA (MG-LDA). Besides the global type of topic, MG-LDA models topics on two levels: global and local. The idea is to have a fixed set of global topics and a dynamic set of local topics, from which the document is sampled. To find the local topics, a document is modeled as a set of sliding windows where each window covers a certain number of adjacent sentences. These windows overlap, causing one particular word to be allowed to be sampled from multiple windows. This also solves the problem of too few co-occurrences: the bags of words are not too small in this case. The set of global topics act in a similar way to the background topic of [58] in Section 3.3.3, increasing the accuracy of the local topics that should represent the sought aspects.

A similar notion is demonstrated in [20] where a distinction is made between global and local topics. Instead of the more complex construction of sliding windows, LDA is simply performed on the sentence level, with the exception that the document topics are modeled in conjunction with the sentence topics. In this way, the sentence topics can model the aspects with all non-relevant words modeled as a document topic.

While finding both global and local topics is useful to get coherent local topics that actually describe aspects, a different option is shown in [42], where LDA is combined with a Hidden Markov Model (HMM) to distinguish between aspect-words and background words. This distinction is drawn by incorporating syntactic dependencies between aspect and sentiment. The same idea can be found in [59], a CRF model discussed in Section 3.3.2, although in [42], it is employed in an unsupervised, generative manner.

Another way of adding syntactic dependencies is shown in [43], where the topic model employs two vocabularies to pick words from. One vocabulary holds the nouns, while the other holds all the words that are dependent on the nouns (e.g., adjectives, adjectival verbs, etc.). These pairs are extracted from the dependency tree as generated by a parser.

In [21], the issue of coverage (cf. [16] in Section 3.1.1) is addressed by estimating the emphasis placed on each aspect by the reviewer. This is done by modeling the overall rating of the product as the weighted sum of the aspect ratings. The inferred weights for the aspect can then be used as a measure of emphasis. However, where [16] returns the reviews which describe a certain aspect most comprehensively based on how much the reviewer is writing about it, [21] determines the emphasis on a certain aspect in a review by its influence on the overall rating. This is an important difference, as the former will show the user reviews that talk much about a certain aspect, even when it is of no consequence to the overall rating, while the latter can output a list of reviews where a certain aspect greatly influences the rating, even when it is barely discussed.

Since LDA models are trained on a per-item basis, a significant number of data points is needed to infer reliable distributions. However, many products on the Web have only a limited number of reviews. Continuing the work on aspect-level sentiment analysis and LDA models, a method to deal with this so-called cold start problem is proposed in [45]. In addition to modeling aspects and sentiment values for products, it also incorporates product categories and the reviewers into the model. By grouping similar products into categories, aspects are associated to product categories instead of the individual products. Then instead of a distribution over all aspects, for each product, only a distribution over the aspects in the product category will have to be derived from the data. Furthermore, this distribution is influenced by the model of the reviewer, which is a distribution over the aspects this reviewer comments on mostly, and with what rating. Hence, a more accurate prediction can be made for products with little or no data.

In [48], a supervised joint aspect and sentiment

model is proposed to determine the helpfulness of reviews on aspect level. The proposed model is a supervised probabilistic graphical model, similar to supervised Latent Dirichlet Allocation. Just like similar LDA models in Section 3.3.3, this model separately and simultaneously models both aspect and sentiment words, to improve the quality of the found aspect topics. While the model is unsupervised with respect to aspect detection, it uses the helpfulness ratings provided for each review as supervision. Unfortunately, because the focus of this work is on the helpfulness prediction, the aspect detection part is not quantitatively evaluated.

3.1.5 Hybrid Methods

Every classification system has its exceptions, and the classification system used in this survey is no different. This section showcases work that falls in more than one of the above categories. When two types of methods are used, they are called hybrid methods and they come in two flavors: serial hybridization, where the output of one phase (e.g., frequency information) forms the input for the next phase (e.g., a classifier or clustering algorithm), and parallel hybridization, where two or more methods are used to find complementary sets of aspects.

Serial hybridization can be found in [49], where Pointwise Mutual Information [60] is used to find possible aspects, which are then fed into a Naïve Bayes classifier to output a set of explicit aspects. Other examples of serial hybridization include [51], where the Dice similarity measure [61] is used to cluster noun phrases that are about the same aspect, and [50] which targets pros and cons to find aspects using frequent nouns and noun phrases, feeding those into an SVM classifier to make the final decision whether it is an aspect or not.

Contrary to the above, a form of parallel hybridization can be found in [52], where a MaxEnt classifier is used to find the frequent aspects, for which there is ample data, and a rule-based method that uses frequency information and syntactic patterns to find the less frequent ones. In this way, available data is used to drive aspect detection, with a rule-based method that acts as back-up for cases where there is not enough data available.

3.2 Sentiment Analysis

The second part of aspect-level sentiment analysis is the actual sentiment analysis, which is the task of assigning a sentiment score to each aspect. The first proposed approaches generally use a dictionary to find the sentiment scores for the individual words followed by an aggregation and/or association step to assign the sentiment of the surrounding words to the aspect itself. The later approaches are all based on machine learning, either supervised or unsupervised. All the approaches that are discussed in this section can be found in Table 2, where their reported performance is also shown.

3.2.1 Dictionary-based

In [26], a sentiment dictionary is obtained by propagating the known sentiment of a few seed words through the WordNet synonym/antonym graph. Only adjectives are considered as sentiment words here. Each adjective in a sentence will be assigned a sentiment class (i.e., positive or negative) from the generated sentiment dictionary. When a negation word appears within a word distance of five words starting from the sentiment word, its polarity is flipped. Then, a sentiment class is determined for each sentence using majority voting. Hence, the same sentiment class is assigned to each aspect within that sentence. However, when the number of positive and negative words is the same, a different procedure is used. In that case, each sentiment bearing adjective is associated with the closest aspect within the sentence, in terms of word distance. Then majority voting is used among all sentiment words that are associated with the same aspect. In this case, having multiple polarities within the same sentence is a possibility.

In contrast to other dictionary methods, [18] uses a set of adjectives provided by Epinions.com, where each adjective is mapped to a certain star rating. The unknown sentiment word, if it is not in this set, is then located in the WordNet synonymy graph. Employing a breadth-first search on the WordNet synonymy graph starting at the adjective with the unknown sentiment with a maximum depth of 5, the two closest adjectives which appear in the rated list of Epinions.com are found. Then, using a distance-weighted nearest-neighbor algorithm, it assigns the weighted average of the ratings of the two nearest neighbors as the estimated rating to the current adjective.

When performing sentiment analysis, some approaches, like the previously discussed [26], compute one sentiment score for each sentence and then associate that sentiment with all the aspects that are mentioned in that sentence. However, this makes it impossible to properly deal with sentences that contain aspects with varying sentiment. A solution is proposed in [62], where all sentences are segmented with each segment being assigned to one of the aspects found in the sentence. Then, using a sentiment lexicon, the polarity of each segment is determined and an aspect-polarity pair is generated that reflects the overall polarity for this aspect within a particular review.

3.2.2 Supervised Machine Learning

While the methods in the previous section all use a dictionary as the main source for information, supervised machine learning methods usually learn many

TABLE 2
Approaches for sentiment analysis

	domain	classes	evaluation task	performance
dictionary-based				
Hu & Liu (2004) [26]	product reviews	binary	sentiment classification	accuracy: 84.2%
Moghaddam & Ester (2010) [18]	product reviews	5-star rating	sentiment classification	Ranking Loss: 0.49
Zhu et al. (2009) [62]	restaurant reviews	ternary	aspect-sentiment extraction	precision: 75.5%
supervised machine learning				
Blair-Goldensohn et al. (2008) [52]	restaurant & hotel reviews	binary	sentiment classification (pos/neg)	precision: 68.0% / 77.2% recall: 90.7% / 86.3%
Yu et al. (2011) [50]	product reviews	binary	sentiment classification	F_1: 71.7%-85.1%
Choi & Cardie (2008) [63]	MPQA corpus [64]	binary	sentiment classification	accuracy: 90.70%
Lu et al. (2011) [20]	restaurant [41] & hotel [40] reviews	5-star rating	sentiment classification	LAE: 0.560 - 0.790
Titov & McDonald (2008) [39]	product, hotel, and restaurant reviews	binary	sentiment classification	Ranking Loss: 0.669
unsupervised machine learning				
Popescu & Etzioni (2005) [49]	product reviews (data from [26])	ternary	sentiment extraction sentiment classification	precision: 76.68% recall: 77.44% precision: 84.8% recall: 89.28%

of their parameters from the data. However, since it is relatively easy to incorporate lexicon information as features into a supervised classifier, many of them employ one or more sentiment lexicons. In [52], the raw score from the sentiment lexicon and some derivative measures (e.g., a measure called purity that reflects the fraction of positive to negative sentiment, thus showing whether sentiment is conflicted or uniform) are used as features for a MaxEnt classifier. When available, the overall star rating of the review is used as an additional signal to find the sentiment of each aspect (cf. [29]).

In [50], the short descriptions in the 'pros' and 'cons' section of a review are mined for sentiment terms. These sentiment terms are found using a dictionary [65], with the location (i.e., either the 'pros' or 'cons' section) denoting their sentiment in that specific context. This information is then used to train a Support Vector Machine (SVM) that is able to classify sentiment terms as positive or negative. Given a free text review, for each aspect, the expression that contains its sentiment is found, which should be within a distance of five steps in the parse tree. Then, the SVM is used to determine the sentiment for that aspect.

While not exactly an aspect-level sentiment analysis method, [63] is still interesting as it performs sentiment analysis on very short expressions, which can be associated to aspects (cf. [62]). Since this method focuses solely on sentiment analysis, the expressions (i.e., short phrases expressing one sentiment on one aspect or entity) are given for this approach. The proposed method is a binary sentiment classifier based on an SVM. But while basic SVM approaches model the text using a simple bag-of-words model, the authors argue that such a model is too simple to represent an expression effectively. To solve this, the authors used the principle of compositional semantics, which states that the meaning of an expression is a function of the meaning of its parts and the syntactic rules by which these are combined. Applying this principle, a two-step process is proposed in which the polarities of the parts are determined first, and then these polarities are combined bottom-up to form the polarity of the expression as a whole. However, instead of using a manually-defined rule set to combine the various parts and their polarities, a learning algorithm is employed to cope with the irregularities and complexities of natural language.

The learning algorithm of the previous approach consists of a compositional inference model using rules incorporated into the SVM update method and a set of hidden variables to encode words being positive, negative, negator, or none of these types. The negator class includes both function-negators and content-negators. While function-negators are only a small set of words like "not" and "never", content-negators are words like "eliminated" and "solve", which also reverse the polarity of their surroundings. As machine learning approaches allow many features, they combine multiple lexicons, adding sentiment information from both the General Inquirer lexicon as well as from the polarity lexicon from Wilson et al. [65]. With some simple heuristics and less sophisticated versions of the proposed method as a baseline, the above solution is evaluated on the MPQA corpus [64]. Experiments show that using compositional inference is more beneficial than using a learning approach, but incorporating both clearly results in the highest accuracy.

Instead of a binary sentiment classifier, as is used in the above two methods [50] [63], a Support Vector Regression model is employed in [20] to find the sentiment score for an aspect. This allows the sentiment score to be modeled as a real number in the zero to five interval, which is reminiscent of the widely used

discrete 5-star rating system.

In [39], a perceptron-based online learning method called PRanking [17], is used to perform the sentiment analysis, given the topic clusters that have been detected by an LDA-like model. The input consists of unigrams, bigrams, and frequent trigrams, plus binary features that describe the LDA clusters. For each sentence, a feature vector $\mathbf{x}$ is constructed consisting of binary features that signal the absence or presence of a certain word-topic-probability combination, with probabilities being grouped into buckets (e.g., 'steak', 'food', and '0.3-0.4'). The PRanking algorithm then takes the inner product of this vector ($\mathbf{x}$) and a vector of learned weights ($\mathbf{w}$) to arrive at a number, which is checked against a set of boundary values that divide the range a score can have into five separate ranges such that each range corresponds to a sentiment value (e.g. one to five). In the training phase, each misclassified instance will trigger an update where both the weights and the boundary values are changed. For example, if an instance is given a sentiment value which is too low, it will both increase weights and decrease threshold values.

3.2.3 Unsupervised Machine Learning

Another option is the use of an unsupervised machine learning method. In [49], each explicit aspect is used to find a potential sentiment phrase by looking for an sentiment phrase in its vicinity, where vicinity is measured using the parsed syntactic dependencies. Each potential sentiment phrase is then examined, and only the ones that show a positive or negative sentiment are retained. The semantic orientation, or polarity, is determined using an unsupervised technique from the computer vision area called relaxation labeling [66]. The task is to assign a polarity label to each sentiment phrase, while adhering to a set of constraints. These constraints arise for example from conjunctions and disjunctions [67]. The final output is a set of sentiment phrases with their most likely polarity label, be it positive or negative.

3.3 Joint Aspect Detection and Sentiment Analysis Methods

All approaches discussed until now either have a method or model dedicated to either aspect detection or sentiment analysis. Since the two problems are not independent, multiple approaches have been proposed that both extract the aspects and determine their sentiment. The main advantage is that combining these two tasks allows one to use sentiment information to find aspects and aspects to find sentiment information. Some methods explicitly model this synergy, while others use it in a more implicit way. We distinguish between syntax-based, supervised machine learning, unsupervised machine learning, and hybrid methods. In Table 3, all approaches discussed in this section are shown, together with their reported performance.

3.3.1 Syntax-Based Methods

Given the observation that it is much easier to find sentiment words than aspect words, syntax-based methods are generally designed to first detect sentiment words, and then by using the grammatical relation between a sentiment word and the aspect it is about, to find the actual aspect. A major advantage of this method is that low-frequency aspects can also be found, as the key factor here is the grammatical relation between the aspect and its sentiment word(s). This is also its greatest shortcoming, since patterns have to be defined that describe the set of possible relations between an aspect and a sentiment word. Unfortunately, a very specific set of relations will miss a lot of aspects leading to high precision, but low recall, while a more general set of relations will yield more aspects but also many more words that are not aspects, leading to low precision, but high recall. Additionally, the extraction of grammatical relations (usually) requires parsing the text, which is both slow and usually not error-free.

An early syntax-based method is [68], where a shallow parser and an extensive set of rules is used to detect aspects and sentiment. The lexicon describes not just the sentiment for a given word, but also gives transfer patterns stating which words are affected by the sentiment. In this way, sentiment originates at a certain word, and is transferred by other words (e.g., verbs) to the aspect word. A good example would be the sentence "The automatic zoom prevents blurry pictures", where negative sentiment originates at 'blurry' and is reversed by the verb 'prevents', transferring the now reversed sentiment to the aspect 'automatic zoom'. Because the described relations are very specific, the result is a typical high-precision low-recall approach that, therefore, works best on large volumes of data.

While most of the previously described approaches focus on product reviews, in [36], an aspect-level sentiment analysis approach is proposed for the movie review domain. This approach employs a lexicon for both the aspect detection and the sentiment analysis part. While the latter is common practice, the former is more of an exception. The intuition behind this is that a lexicon can capture all the domain specific cues for aspects. For example, this aspect lexicon includes a list of names of people involved in the movie that is under review. Dependency patterns that link the aspect and the sentiment word are used to find aspect-sentiment pairs. However, the described relations only cover the most frequent relations, so less frequent ones are missed.

TABLE 3
Approaches for joint aspect detection and sentiment analysis

	domain	classes	evaluation task	performance
syntax-based				
Nasukawa & Yi (2003) [68]	general & camera reviews	binary	combined (general/camera)	precision: 94.3% / 94.5% recall: 28.6% / 24%
Zhuang et al. (2006) [36]	movie reviews	binary	aspect-sentiment pair mining	F_1: 52.9%
supervised machine learning				
Kobayashi et al. (2006) [69]	product reviews	binary	sentiment extraction	precision: 67.7% recall: 50.7%
			aspect-sentiment pair mining	precision: 76.6% recall: 75.1%
			sentiment classification	precision: 82.2% recall: 66.2%
Li et al. (2010) [59]	product & movie reviews	binary	combined (movies/products)	precision: 82.6% / 86.6% recall: 76.2% / 69.3%
Marcheggiani et al. (2014) [19]	hotel reviews (annotated subset of [40])	ternary	aspect detection sentiment classification	F_1: 48.5% MAE^M: 0.5
Jin et al. (2009) [70]	camera reviews	binary	aspect extraction sentiment sentence extraction sentiment classification	F_1: 78.8% - 82.7% F_1: 84.81% - 88.52% F_1: 70.59% - 77.15%
Zirn et al. (2011) [71]	product reviews	binary	combined (pos/neg)	precision: 66.38%/72.02% recall: 72.94%/65.34%
unsupervised machine learning				
Mei et al. (2007) [58]	weblogs	binary	sentiment model (pos / neg)	KL-divergence: 21 / 19
Titov & McDonald (2008) [72]	hotel reviews	5-star rating	combined	avg. prec.: 74.5% - 87.6%
Moghaddam & Ester (2011) [73]	product reviews	5-star rating	aspect detection sentiment classification	Rand Index: 0.83 Rand Index: 0.73
Jo & Oh (2011) [74]	restaurant & product reviews	binary	sentiment classification	accuracy: 84% - 86%
Wang et al. (2011) [21]	mp3 player & hotel [40] reviews	5-star rating	aspect rating prediction	MSE: 1.234 nDCG: 0.901
Sauper & Barzilay (2013) [75]	restaurant reviews	binary	aspect cluster prediction	precision: 74.3% recall: 86.3%
			sentiment classification	accuracy: 82.5%
	medical summaries		aspect cluster prediction	precision: 89.1% recall: 93.4%
hybrid machine learning				
Zhao et al. (2010) [76]	restaurant [77] & hotel [44] reviews	ternary + 'conflicted'	aspect identification sentiment identification	avg. F_1: 70.5% precision @ 5: 82.5% precision @ 10: 70.0%
Mukherjee & Liu (2012) [78]	product reviews	binary	sentiment classification	precision: 78% recall: 73%

3.3.2 Supervised Machine Learning

An evident problem is that in general, machine learning methods excel in classifying instances in a given number of classes. Since the number of possible aspects and the different words that can represent an aspect is practically unbounded, a default classification algorithm cannot be applied in a straightforward manner. In [69], both aspect detection and sentiment classification are cast as a binary classification problem. First, using a lexicon, all prospective aspect and sentiment words are tagged. Then, the problem of which aspect belongs to which sentiment word is solved using a binary classification tournament model. Each round of the tournament, two aspects are compared and the one that best matches the sentiment word proceeds to the next round. In this way, no direct relation between the aspect and sentiment is needed. The drawback is that no additional aspects can be found by exploiting this relation, but an advantage is that this method can effectively deal with ellipsis, a linguistic phenomenon where the aspect is not linked to the sentiment because it is either implicit or referred to using a co-reference. According to the authors, as much as 30% of the sentences feature ellipsis.

To address the issue of long-range dependencies, [59] encodes both syntactic dependencies between words and conjunctions between words into a CRF model. By introducing more dependencies between the hidden nodes in the CRF model, words that are not directly adjacent in the linear chain CRF, can now influence each other. Sentiment values and their targets are linked simply by minimizing the word distance and are extracted simultaneously. The model is then used to generate a list of sentiment-entity pairs as a summary of the set of texts, which are product and movie reviews in this case, grouped as positive or negative.

A strong limitation of the previous work is that each sentence is assumed to have only one aspect. In [19], a CRF model is proposed that is able to

deal with multiple aspects per sentence. Furthermore, when multiple aspects are mentioned in the same sentence, it is likely that they influence each other via certain discourse elements, which has an effect on the sentiment score for each aspect. Therefore, the model explicitly incorporates the relations between aspect-specific sentiments within one sentence. Last, the overall score of the review, which is often supplied by the users themselves, is taken into account as well. To do that, a hierarchical model is proposed that simultaneously predicts the overall rating and the aspect ratings. This new model has an additional variable for the overall sentiment score, and pairwise factors that model the influence between the overall sentiment score and each aspect's sentiment score. A random subset of 369 hotel reviews from the TripAdvisor data set [40] is manually annotated for aspects to train and test the model.

An example of a method based on a lexicalized HMM is [70]. With HMM's, the context of a word can easily be taken into consideration by using n-grams. However, simply using higher n-grams (e.g., bigrams, trigrams, etc.) poses some problems. Because a lot of these n-grams are not likely to appear in the training corpus, their values have to be guessed instead of counted. Furthermore, computational complexity increases exponentially when using higher n-grams. This is the reason that in [70] only unigrams are used. While this prevents the above mentioned problems, it also deprives the model of any context-sensitivity. To account for it, the part-of-speech of a word is also modeled, and in a way that makes it dependent on both the previous and the next part-of-speech tag, thereby introducing some form of context-awareness. A bootstrapping approach is proposed to make the model self-learn a lot of training examples, mitigating the dependence on labeled training data to some extent. The additional examples learned in this way proved to be beneficial when evaluating this approach, improving F_1-score for both aspect detection and sentiment classification.

A Markov logic chain is employed as the main learning method in [71]. Within the Markov chain, multiple lexicons are incorporated, as well as discourse relations. The latter are acquired using the HILDA [79] discourse parser which returns a coarse-grained set of discourse segments as defined in [80], which are based on the Rhetorical Structure Theory [81]. Since sentiment classification is done on the level of discourse segments, it is assumed each segment only expresses one sentiment, which is almost always the case. Entities, however, are not extracted in this method. The proposed classification in [71] is binary, which, according to the authors, results in problems with some segments that have no clear polarity. Their findings concerning the use of discourse elements were that using general structures that can be found in the text systematically improves the results. The fact that a certain discourse relation describes a contrasting relation was encoded specifically, as it was expected to correlate with the reversing of polarity of the various segments it connects to. However, this correlation turned out to be not as strong as was expected beforehand. This means, according to the authors, that the classical discourse relations might not be the best choice to represent the general structure of the text when performing sentiment analysis. Nevertheless, the same authors believe that focusing on cue words to find discourse connectives in order to predict polarity reversals might still be worth investigating.

3.3.3 Unsupervised Machine Learning

The class of unsupervised machine learning approaches may be especially interesting, since these models are able to perform both aspect detection and sentiment analysis without the use of labeled training data. The first topic mixture model [58] is based on probabilistic Latent Semantic Indexing (PLSI) [55], a model similar to LDA, that is however more prone to overfitting and is not as statistically sound as LDA. In [58], not only topics that correspond to aspects are modeled, but also a topic for all background words, causing the retrieved topics to better correspond to the actual aspects. Furthermore, the topics that correspond to aspects are again mixtures of sentiment topics. In this way, the end result is that both aspects and their sentiment are determined simultaneously with the same model. Leveraging a sentiment lexicon to better estimate the sentiment priors increases the accuracy of the sentiment classification.

In [72], Titov and McDonald extend the model they propose in [39] by including sentiment analysis for the found aspects. An additional observed variable is now added to the model, namely the aspect ratings provided by the author of the review. With the assumption that the text is predictive of the rating provided by the author, this information can be leveraged to improve the predictions of the model. A strong point is that the model does not rely on this information being present, but when present, it is used to improve the model's predictions. Besides utilizing the available aspect ratings, the model can extract other aspects from the text as well, and assign a sentiment score to them. While at least a certain amount of provided aspect ratings is needed for this model to truly benefit from them, perhaps the biggest advantage is that the found aspects can be linked to actual aspects in the text. As mentioned earlier, generative models produce unlabeled clusters that are not associated with any particular aspect. This problem is solved by incorporating these aspect ratings into the LDA model, providing a link between the words in the document and the concrete aspects as annotated by the reviewer. Last, when put to the test against a

MaxEnt classifier, a supervised method, the proposed method performed only slightly worse.

The main improvement of [73], compared to previous topic models is that the sentiment class of an aspect is explicitly linked to the aspect itself. This makes the sentiment analysis more context-aware: in this way, a word that is positive for one aspect can be negative for another. The latter is generally true for models that couple the sentiment nodes to the aspect nodes in the graphical model, and this same idea is demonstrated in both [21] and [74].

In [74], aspects are detected as topics by constraining the model to only one aspect-sentiment combination per sentence. By assuming that each sentence is about only one aspect and conveys only one sentiment, the model is able to find meaningful topics. This is a relatively simple solution compared to for example the sliding windows technique [39] or injecting syntactic knowledge into the topic model [42]. Evaluation of the constructed topics revealed another interesting fact: in one particular case there were three topics that conveyed negative sentiment for the same aspect. While this may not seem ideal at first (i.e., one unique topic per aspect-sentiment combination is more logical), close inspection revealed that the three topics revealed three distinct reasons why the reviewers were negative about that aspect (i.e., the screen was too small, the screen was too reflective, and the screen was easily covered with fingerprints or dirt). This level of detail goes further than regular aspect-level sentiment analysis, providing not only the sentiment of the reviewers, but also the arguments and reasons why that sentiment is associated to that aspect.

In [75], a probabilistic model is presented that performs joint aspect detection and sentiment analysis for the restaurant reviews domain and aspect detection alone for the medical domain. For the restaurant domain, it models the aspects in such a way that they are dependent on the entity (i.e., the restaurant), instead of having a global word distribution for aspects like previous models. This allows the model to have different aspects for different kind of restaurants. For example, a steak house has different aspects than an Italian ice cream place and while the sentiment word distribution is global (i.e., the same sentiment words are used for all types of restaurants), a separate distribution that is different for each restaurant is used to model the link between aspects and sentiment words. Furthermore, an HMM-based transition function is employed to model the fact that aspects and sentiment words often appear in a certain order. Last, a background word distribution is determined on a global level to get rid of words that are irrelevant. A variant of the model is used to process dictated patient summaries. Since the set of relevant aspects is expected to be shared across all summaries, the aspects are modeled as global word distribution. The previous method operates in an unsupervised fashion, requiring only a set of sentiment seed words to bias the sentiment topics into a specific polarity. Furthermore, the proposed model admits an efficient inference procedure.

3.3.4 Hybrid Machine Learning

While LDA is designed to work with plain text, the above methods have shown that the right preprocessing can significantly improve the results of the generative model. This can be extended a bit further by already optimizing some of the input for the topic model by using a supervised discriminative method. Both methods presented in this section feature a MaxEnt classifier that optimizes some of the input for the LDA model.

The first method [76] uses a MaxEnt component to enrich the LDA model with part-of-speech information. In this way, the generative model can better distinguish between sentiment words, aspect words, and background words. The MaxEnt classifier is trained using a relatively small set of labeled training data, and the learned weights are now input for a hidden node in the topic model. This is done before training the LDA model, so while training the LDA model, the weights of the MaxEnt classifier remain fixed.

The second method [78] that combines an LDA model with a MaxEnt classifier, uses the MaxEnt classifier to optimize the word priors that influence the generative process of drawing words. Again, part-of-speech information is a major feature for the MaxEnt component. The fact that external information can be integrated into the generative process of an LDA model makes it a very powerful and popular method for aspect-level sentiment analysis.

4 Related Issues

While finding aspects and determining their sentiment value is the core of aspect-level sentiment analysis, there are more issues that play a role in developing an effective tool for aspect-level sentiment analysis. This section discusses some of these related issues. First, a set of sub-problems will be discussed, including how to deal with comparative opinions, conditional sentences, and negations and other modifiers. Then, a short discussion on aggregation of sentiment scores is given, followed by a concise exposition on presentation of aspect-level sentiment analysis results.

4.1 Sub-problems

Processing natural language in general, and performing aspect-level sentiment analysis, specifically, is a very complex endeavor. Therefore, it has been proposed, for example in [82], that instead of focusing on a one-size-fits-all solution, researchers should focus on the many sub-problems. By solving enough of the sub-problems, the problem as a whole can

eventually be solved as well. This line of thought has given rise to work specifically targeting a certain sub-problem in sentiment analysis, which is discussed below. The presented approaches are not solutions for aspect-level sentiment analysis and are therefore not in the tables together with the previously discussed approaches. However, when aspect-level sentiment analysis methods take the issues presented below into account (and some do to some extent), performance will increase.

4.1.1 Comparative Opinions

In comparative sentences, one entity or aspect is usually compared with another entity or aspect by preferring one over the other. Detecting comparative sentences and finding the entities and aspects that are compared, as well as the comparative words themselves is very useful [83]. However, for sentiment analysis, one really needs to know which entity or aspect is preferred, a problem that is discussed in [84].

First, various categories of comparative sentences are defined and, for each category, it is shown how to process them. When possible, a comparator is reduced to its base form, and its sentiment is found using the sentiment word list generated from WordNet [26]. The comparators whose polarity cannot be determined in this way are labeled as context-dependent and are processed differently. For that, information in the pros and cons section is leveraged to compute an asymmetric version of the Pointwise Mutual Information association score between the comparative words and the words in the pros and cons. A set of rules then essentially combines the information about the entities, comparative words, and aspects being compared into one coherent outcome: either a positive or negative sentiment about the preferred entity.

A remaining problem in [84] is that when something is more positive than something else, the first is assumed to have a positive sentiment. This is not always the case. Also problematic is the negation of comparators, as stated by the authors themselves. Their example of "not longer" not necessarily being the same as "shorter" is illustrative. While the proposed method currently perceives the second entity as the preferred one when encountering negations, the authors admit that it could also be the case that the user did not specify any preference.

4.1.2 Conditional Sentences

As discussed in the previous section, conditional sentences pose a problem in that it is hard to determine whether they actually express some sentiment on something or not. In [82], an approach dedicated to conditional sentences was proposed, which can be seen as an extension of the existing line of research based on [26]. First, the various types of conditionals were grouped into four categories, each with part-of-speech patterns for both the condition and the consequent in that category. Around 95% of the targeted sentences is covered by these patterns. The sentences found are then classified as either positive, negative, or neutral with respect to some topic in that sentence. For this study, the topic is assumed to be known beforehand. In contrast to previously described research, the authors chose to use an SVM to classify these sentences as having either a positive or negative polarity.

Features used for the SVM are the basic ones like sentiment words and part-of-speech information, but also some common phrases and a list of words that imply the lack of sentiment. Also covered are negations by adding a list of negation keywords. This is, however, still based on a simple word distance metric. Other notable features are the fact whether the topic is in the conditional or the consequent and the length of both the condition and consequent phrases. Last, the sentiment words were weighted according to the inverse of their distance to the topic.

Multiple ways of training were proposed in [82], but using the whole sentence instead of only the conditional or consequent part turned out to be the most successful. Interestingly, while the whole-sentence classifier gave the best results, the consequent-only classifier gave much better results than the conditional-only classifier, even approaching the results of the whole-sentence classifier, suggesting that most useful information to classify conditionals is in the consequent and not in the conditional part. The classifier was trained on a set of product reviews which were manually annotated and tested on both a binary and a ternary classification problem.

For the binary classification, the consequent-only classifier and the whole-sentence classifier yielded a similar performance while for the ternary classification, the whole-sentence approach performed clearly better. According to the authors, this signifies that to classify something as neutral, information from both the conditional and the consequent are needed. The best result the authors reported is an accuracy of 75.6% for the binary classification and 66.0% for the ternary classification. Unfortunately, no baseline was defined to compare these results against.

4.1.3 Negations and Other Modifiers

From amongst the set of modifiers that change the polarity or strength of some sentiment, negations are implemented most. This comes to no surprise given the effect negations can have on the sentiment of an aspect, sentence, or document. A theoretical discussion by Polyani and Zaenen [85] proposes some foundational considerations when dealing with these contextual valence shifters as they are sometimes called. The authors distinguish between sentence-based contextual valence shifters and discourse-based ones.

Negations and intensifiers, which belong to the sentence-based group, are mostly single words influencing the polarity of words that are within their scope. Negations flip the polarity of a sentiment, while intensifiers either increase or decrease the sentiment value. Other sentence-based contextual valence shifters are: modals, where a context of possibility or necessity is created as opposed to real events (e.g., "if she is such a brilliant person, she must be socially incapable."); presuppositional items which represent certain expectation that are met or not (e.g., "this is barely sufficient"); and irony in which overly positive or negative phrases are turned on themselves to create a sentence with the opposite valence or polarity (e.g., "the solid and trustworthy bank turned to robbing their own customers").

The category of discourse-based contextual valence shifters is more complex in nature. While one group, the discourse connectors, are linked to some particular words, all other categories are much harder to identify. We will therefore only briefly discuss these discourse connectors, and refer the interested reader to [85] for more categories. Discourse connectors are words that connect two or more phrases in such a way that the combination is different in terms of sentiment than simply the sum of its parts. An example to illustrate this is "while he is grumpy each day, he is not a bad person", where we can see that the connector 'while' mitigates the effects of 'grumpy', resulting in an overall positive sentence.

An implementation of the above framework was described in [86], where many of the ideas of Polyani and Zaenen are encoded in rules. The resulting pipeline, which also included a part-of-speech tagger and a parser, was evaluated to analyze where errors do occur. The results are rather interesting, as about two-thirds of the errors occur before the valence shifting module. Large contributions to errors are made by the parser and the tagger (around 14% each) and the lack of a word sense disambiguation module (25%). Errors made by the valence shifter module can roughly be attributed to three reasons: either the polarity reading was ambiguous (10%), more world knowledge was required (19%), or the polarity was modulated by phenomena more closely related to pragmatics than semantics (5%).

While Polyani and Zaenen did not really discuss the scope of a negation, this is actually a very important topic. Most approaches to sentiment analysis have at least some handling of negations, but they usually employ only a simple word distance metric to determine which words are affected by a negation keyword (cf. [87] for a comparison of different word distances). In [88], the concept of the scope of a negation term is further developed. For each negation term, its scope is found by using a combination of parse tree information and a set of rules. The general idea is to use the parse tree to find the least common ancestor of the negation word and the word immediately following it in the sentence. Then all leaves descending from that ancestor that are to the right of the negation term are in the scope of the negation. This scope is then further delimited and updated by the set of rules to cover some exceptions to this general rule.

When looking at informal texts, such as microblog posts, additional modifiers need to be taken into account [89]. Lexical variants that intensify the expressed sentiment include the use of repeated punctuation with exclamation marks and using repeated characters inside a word (e.g., 'haaaaaappy'). Other sources of sentiment that are employed in informal texts are emoticons, for which a custom list of emoticons with their sentiment score is usually needed.

4.2 Aggregation

Several of the discussed approaches aggregate sentiment over aspects, usually to show an aspect-based sentiment summary. Most methods aggregate sentiment by simply averaging or taking a majority vote. In contrast, methods that employ topic models, for example [72], aggregate naturally over the whole corpus, thereby computing sentiment for each topic or aspect based on all the reviews. A different approach is shown [40], where the topic model does not return the aggregated aspect ratings, but instead presents the aspect ratings for each individual review, as well as the relative weight placed on that aspect by the reviewer. The authors discuss that this enables advanced methods of aggregation, where aspect ratings can be weighted according to the emphasis placed on it by each reviewer.

In [90], multiple methods for aggregating sentiment scores are investigated. Even though this work focuses on combining sentence-level sentiment scores into a document-level sentiment score, the ideas can be naturally translated into the domain of aspect-level sentiment analysis. Next to a series of heuristic methods, a formally defined method for aggregation based on the Dempster-Shafer Theory of Evidence [91] is proposed. This is a theory of uncertainty that can be used to quantify the amount of evidence a certain source contributes to some proposition. In this case, the sources of evidence are the sentence sentiment scores, and the proposition to which these sources of evidence contribute is the final document-level sentiment score.

The following methods of aggregation are tested: randomly picking a sentence sentiment as the document sentiment, simply averaging all sentence sentiment scores, taking the absolute maximum score (e.g., when the strongest positive sentence is $+5$ and the strongest negative sentence is -4, the overall sentiment will be $+5$), summing the two maximum scores (e.g., in the previous example, summing $+5$ and -4 would result in a $+1$ document-level sentiment),

scaled rate which is the fraction of positive sentiment words out of all sentiment words, and the discussed Dempster-Shafer method. As shown in [90], the proposed method clearly outperforms all heuristics. It is argued that this is caused by the fact that the Dempster-Shafer method takes all pieces of evidence into account, and the fact that it considers maximal agreements among the pieces of evidence. Of interest is the fact that this method is tested on two data sets that have also been used for already discussed methods that perform aspect-level sentiment analysis (cf. Tables 1, 2, and 3). Hence, methods for aspect-level sentiment analysis should be able to benefit from this research.

4.3 Presentation

As a final step in the process of aspect-level sentiment analysis, the results should be presented to the user. This can be done in several ways, the first of which is simply showing the numbers. In this case, for a certain product, a list of detected aspects is shown, together with the aggregated sentiment scores for each aspect. One can also imagine a table with the scores for multiple products in order to easily compare them.

In [28], a visual format is advocated that shows bars that denote the sentiment scores. Clicking the bar would show more details, including relevant snippets of reviews. In this way, a user can quickly inspect the traits of several products and compare them, without getting overwhelmed by a table full of numbers. When the timestamp of each review is available, a timeline [92] could also be generated to show the change in sentiment over time. This is important for services, which can change over time, or product characteristics which may only show after prolonged use.

Another possibility is to generate a summary of all the analyzed reviews. When done right, this will produce a readable review that incorporates all the available information spread over all reviews. In [93], an ontology is used to organize all the aspects into aspect categories and all sentences that express sentiment on an aspect are linked to the aspects in the ontology as well. Two methods for summary generation are tested: the first is to select representative sentences from the ontology, the second is to generate sentences with a language generator based on the aspects and their known sentiment scores. While the sentence selection method yields more variation in the language being used in the summary as well as more details, the sentence generation provides a better sentiment overview of the product. A variation of this method is contrastive summarization [94], where the summary consists of pairs of sentences that express opposing sentiment on the same aspect.

5 Conclusions

From the overview of the state-of-the-art in aspect-level sentiment analysis presented in this survey, it is clear that the field is transcending its early stages. While in some cases, a holistic approach is presented that is able to jointly perform aspect detection and sentiment analysis, in others dedicated algorithms for each of those two tasks are provided. Most approaches that are described in this survey are using machine learning to model language, which is not surprising given the fact that language is a non-random, very complex phenomenon for which a lot of data is available. The latter is especially true for unsupervised models, which are very well represented in this survey.

We would like to stress that transparency and standardization is needed in terms of evaluation methodology and data sets in order to draw firm conclusions about the current state-of-the-art. Benchmark initiatives like SemEval [13], [14] or GERBIL [15] that provide a controlled testing environment are a shining example of how this can be achieved.

When considering the future of aspect-level sentiment analysis, we foresee a move from traditional word-based approaches, towards semantically rich concept-centric aspect-level sentiment analysis [95]. For example, in "This phone doesn't fit in my pocket", it is feasible to determine that the discussed aspect is the size of the phone. However, the negative sentiment conveyed by this sentence, related to the fact that phones are supposed to fit in one's pocket, seems extremely hard to find for word-based methods. Related to this problem, pointing to the need for reasoning functionality, is the still open research question of irony. In [96], a conceptual model is presented that explicitly models expectations, which is necessary to effectively detect irony. This is also a step away from the traditional word-based approach towards a semantic model for natural language processing. While concept-centric, semantic approaches have only recently begun to emerge (e.g., ontologies are being used to improve aspect detection [97]), they should be up to this challenge, since semantic approaches naturally integrate common sense knowledge, general world knowledge, and domain knowledge.

Combining concept-centric approaches with the power of machine learning will give rise to algorithms that are able to reason with language and concepts at a whole new level. This will allow future applications to deal with complex language structures and to leverage the available human-created knowledge bases. Additionally, this will enable many application domains to benefit from the knowledge obtained from aspect-level sentiment analysis.

ACKNOWLEDGMENTS

The authors of this paper are supported by the Dutch national program COMMIT. We would like to thank the reviewers for their invaluable insights. Furthermore, we are grateful for the constructive comments given by Franciska de Jong and Rommert Dekker.

REFERENCES

[1] B. Bickart and R. M. Schindler, "Internet Forums as Influential Sources of Consumer Information," *Journal of Interactive Marketing*, vol. 15, no. 3, pp. 31–40, 2001.

[2] Ellen van Kleef and Hans C.M. van Trijp and Pieternel Luning, "Consumer Research in the Early Stages of New Product Development: a Critical Review of Methods and Techniques," *Food Quality and Preference*, vol. 16, no. 3, pp. 181–201, 2005.

[3] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] Y. Chen and J. Xie, "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science*, vol. 54, no. 3, pp. 477–491, 2008.

[5] R. E. Goldsmith and D. Horowitz, "Measuring Motivations for Online Opinion Seeking," *Journal of Interactive Advertising*, vol. 6, no. 2, pp. 3–14, 2006.

[6] I. Arnold and E. Vrugt, "Fundamental Uncertainty and Stock Market Volatility," *Applied Financial Economics*, vol. 18, no. 17, pp. 1425–1440, 2008.

[7] M. Tsytsarau and T. Palpanas, "Survey on Mining Subjective Data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.

[8] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2012, vol. 16.

[9] *Collins English Dictionary Complete and Unabridged*. HarperCollins Publishers, June 2015, "Opinion". [Online]. Available: http://www.thefreedictionary.com

[10] S.-M. Kim and E. Hovy, "Determining the Sentiment of Opinions," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Association for Computational Linguistics, 2004.

[11] R. Plutchik, *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, 1980.

[12] H. Tang, S. Tan, and X. Cheng, "A Survey on Sentiment Detection of Reviews," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10760–10773, 2009.

[13] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, 2014, pp. 27–35.

[14] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 Task 12: Aspect Based Sentiment Analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, 2015, pp. 486–495.

[15] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann, "GERBIL: General Entity Annotator Benchmarking Framework," in *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*. ACM, 2015, pp. 1133–1143.

[16] C. Long, J. Zhang, and X. Zhut, "A Review Selection Approach for Accurate Feature Rating Estimation," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. ACL, 2010, pp. 766–774.

[17] K. Crammer and Y. Singer, "Pranking with Ranking," in *Advances in Neural Information Processing Systems 14 (NIKS 2001)*. MIT Press, 2001, pp. 641–647.

[18] S. Moghaddam and M. Ester, "Opinion Digger: an Unsupervised Opinion Miner from Unstructured Product Reviews," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*. ACM, 2010, pp. 1825–1828.

[19] D. Marcheggiani, O. Täckström, A. Esuli, and F. Sebastiani, "Hierarchical Multi-label Conditional Random Fields for Aspect-Oriented Opinion Mining," in *Proceedings of the 36th European Conference on Information Retrieval (ECIR 2014)*. Springer, 2014, pp. 273–285.

[20] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-Aspect Sentiment Analysis with Topic Models," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*. IEEE, 2011, pp. 81–88.

[21] H. Wang, Y. Lu, and C. Zhai, "Latent Aspect Rating Analysis without Aspect Keyword Supervision," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011)*. ACM, 2011, pp. 618–626.

[22] K. Järvelin and J. Kekäläinen, "Cumulated Gain-based Evaluation of IR Techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[24] S. Moghaddam and M. Ester, "Tutorial at WWW 2013: 'Opinion Mining in Online Reviews: Recent Trends'," 2013. [Online]. Available: http://www.cs.sfu.ca/ ester/papers/WWW2013.Tutorial.Final.pdf

[25] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," in *Proceedings of the 19th National Conference on Artifical Intelligence (AAAI 2004)*. AAAI, 2004, pp. 755–760.

[26] ——, "Mining and Summarizing Customer Reviews," in *Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*. ACM, 2004, pp. 168–177.

[27] Z. Hai, K. Chang, and J.-j. Kim, "Implicit Feature Identification via Co-occurrence Association Rule Mining," in *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text processing (CICLing 2011)*. Springer, 2011, vol. 6608, pp. 393–404.

[28] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*. ACM, 2005, pp. 342–351.

[29] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red Opal: Product-Feature Scoring from Reviews," in *Proceedings of the 8th ACM Conference on Electronic Commerce (EC 2007)*. ACM, 2007, pp. 182–191.

[30] Z. Li, M. Zhang, S. Ma, B. Zhou, and Y. Sun, "Automatic Extraction for Product Feature Words from Comments on the Web," in *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology (AIRS 2009)*. Springer, 2009, pp. 112–123.

[31] Y. Zhao, B. Qin, S. Hu, and T. Liu, "Generalizing Syntactic Structures for Product Attribute Candidate Extraction," in *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies 2010 (HLT-NAACL 2010)*. ACL, 2010, pp. 377–380.

[32] Jun Zhao and Hongbo Xu and Xuanjing Huang and Songbo Tan and Kang Liu and Qi Zhang, "Overview of Chinese Opinion Analysis Evaluation 2008," in *Proceedings of the 1st Chinese Opinion Analysis Evaluation (COAE 2008)*, 2008, pp. 1–21.

[33] G. Qiu, B. Liu, J. Bu, and C. Chen, "Expanding Domain Sentiment Lexicon Through Double Propagation," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence (IJCAI 2009)*. Morgan Kaufmann Publishers Inc., 2009, pp. 1199–1204.

[34] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. ACL, 2010, pp. 1462–1470.

[35] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields," in *Proceedings of the 2010 Conference on Empirical Meth-*

ods in Natural Language Processing (EMNLP 2010). ACL, 2010, pp. 1035–1045.

[36] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie Review Mining and Summarization," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006)*, vol. 6, no. 11. ACM, 2006, pp. 43–50.

[37] C. Toprak, N. Jakob, and I. Gurevych, "Sentence and Expression Level Annotation of Opinions in User-Generated Discourse," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. ACL, 2010, pp. 575–584.

[38] J. S. Kessler and N. Nicolov, "Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations," in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*. AAAI, 2009, pp. 90–97.

[39] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-Grain Topic Models," in *Proceedings of the 17th International Conference on World Wide Web (WWW 2009)*. ACM, 2008, pp. 111–120.

[40] H. Wang, Y. Lu, and C. Zhai, "Latent Aspect Rating Analysis on Review Text Data: a Rating Regression Approach," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and Data Mining (KDD 2010)*. ACM, 2010, pp. 783–792.

[41] G. Ganu, N. Elhadad, and A. Marian, "Beyond the Stars: Improving Ratig Predictions using Review Content," in *Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009)*, 2009.

[42] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu, "Exploiting Coherence for the Simultaneous Discovery of Latent Facets and Associated Sentiments," in *SIAM International Conference on Data Mining 2011 (SDM 2011)*. SIAM, 2011, pp. 498–509.

[43] T.-J. Zhan and C.-H. Li, "Semantic Dependent Word Pairs Generative Model for Fine-Grained Product Feature Mining," in *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2011)*. Springer, 2011, pp. 460–475.

[44] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet Rating of Product Reviews," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*. Springer, 2009, pp. 461–472.

[45] S. Moghaddam and M. Ester, "The FLDA Model for Aspect-based Opinion Mining: Addressing the Cold Start Problem," in *Proceedings of the 22nd International Conference on World Wide Web (WWW 2013)*. ACM, 2013, pp. 909–918.

[46] N. Jindal and B. Liu, "Opinion Spam and Analysis," in *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM 2008)*. ACM, 2008, pp. 219–230.

[47] S. Moghaddam and M. Ester, "On the Design of LDA Models for Aspect-Based Opinion Mining," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*. ACM, 2012, pp. 803–812.

[48] Z. Hai, G. Cong, K. Chang, W. Liu, and P. Cheng, "Coarse-to-fine Review Selection via Supervised Joint Aspect and Sentiment Model," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*. ACM, 2014, pp. 617–626.

[49] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," in *Proceedings of the Conference on Human Language Technology and Conference on Empirical Methods in Natural Language Processing 2005 (HLT/EMNLP 2005)*. ACL, 2005, pp. 339–346.

[50] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*. ACL, 2011, pp. 1496–1505.

[51] S. Raju, P. Pingali, and V. Varma, "An Unsupervised Approach to Product Attribute Extraction," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*. Springer, 2009, pp. 796–800.

[52] S. Blair-Goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar, "Building a Sentiment Summarizer for Local Service Reviews," in *Proceedings of WWW 2008 Workshop on NLP in the Information Explosion Era (NLPIX 2008)*. ACM, 2008.

[53] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," *Computational Linguistics*, vol. 37, no. 1, pp. 9–27, 2011.

[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[55] T. Hofmann, "Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization," in *Advances in Neural Information Processing Systems (NIPS 2000*. MIT Press, 2000, pp. 914–920.

[56] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*. ACM, 2011, pp. 775–784.

[57] D. M. Blei and P. J. Moreno, "Topic Segmentation with an Aspect Hidden Markov Model," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*. ACM, 2001, pp. 343–348.

[58] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs," in *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*. ACM, 2007, pp. 171–180.

[59] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu, "Structure-Aware Review Mining and Summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. ACL, 2010, pp. 653–661.

[60] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[61] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[62] J. Zhu, H. Wang, B. K. Tsou, and M. Zhu, "Multi-Aspect Opinion Polling from Textual Reviews," in *Proceedings of the 18th ACM Conference on Information and knowledge management (CIKM 2009)*. ACM, 2009, pp. 1799–1802.

[63] Y. Choi and C. Cardie, "Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 2008, pp. 793–801.

[64] J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, vol. 39, no. 2, pp. 165–210, 2005.

[65] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings of the Conference on Human Language Technology and Conference on Empirical Methods in Natural Language Processing 2005 (HLT/EMNLP 2005)*. ACL, 2005, pp. 347–354.

[66] R. A. Hummel and S. W. Zucker, "On the Foundations of Relaxation Labeling Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 3, pp. 267 –287, 1983.

[67] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL 1997)*. Morgan Kaufman Publishers / ACL, 1997, pp. 174–181.

[68] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing," in *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003)*. ACM, 2003, pp. 70–77.

[69] N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto, "Opinion mining on the web by extracting subject-aspect-evaluation relations," in *Proceedings of the AAAI Spring Symposium 2006: Computational Approaches to Analyzing Weblogs (AAAI-SS 2006)*. AAAI, 2006, pp. 86–91.

[70] W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: a Novel Machine Learning System for Web Opinion Mining and Extraction," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*. ACM, 2009, pp. 1195–1204.

[71] C. Zirn, M. Niepert, H. Stuckenschmidth, and M. Strube, "Fine-Grained Sentiment Analysis with Structural Features,"

in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*. ACL, 2011, pp. 336–344.

[72] I. Titov and R. McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2008)*. ACL, 2008, pp. 308–316.

[73] S. Moghaddam and M. Ester, "ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. ACM, 2011, pp. 665–674.

[74] Y. Jo and A. H. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," in *Proceedings of the Forth International Conference on Web Search and Web Data Mining (WSDM 2011)*. ACM, 2011, pp. 815–824.

[75] C. Sauper and R. Barzilay, "Automatic Aggregation by Joint Modeling of Aspects and Values," *Journal of Artificial Intelligence Research*, vol. 46, no. 1, pp. 89–127, 2013.

[76] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. ACL, 2010, pp. 56–65.

[77] S. Brody and N. Elhadad, "An Unsupervised Aspect-Sentiment Model for Online Reviews," in *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies 2010 (HLT-NAACL 2010)*. ACL, 2010, pp. 804–812.

[78] A. Mukherjee and B. Liu, "Modeling Review Comments," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. ACL, 2012, pp. 320–329.

[79] D. A. duVerle and H. Prendinger, "A Novel Discourse Parser Based on Support Vector Machine Classification," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL 2009)*, 2009, pp. 665–673.

[80] R. Soricut and D. Marcu, "Sentence Level Discourse Parsing Using Syntactic and Lexical Information," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL 2003)*. ACL, 2003, pp. 149–156.

[81] W. C. Mann and S. A. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization," *Text*, vol. 8, no. 3, pp. 243–281, 1998.

[82] R. Narayanan, B. Liu, and A. Choudhary, "Sentiment Analysis of Conditional Sentences," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*. ACL, 2009, pp. 180–189.

[83] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*. ACM, 2006, pp. 244–251.

[84] M. Ganapathibhotla and B. Liu, "Mining Opinions in Comparative Sentences," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*. ACL, 2008, pp. 241–248.

[85] L. Polanyi and A. Zaenen, "Contextual Valence Shifters," in *Computing Attitude and Affect in Text: Theory and Applications*, ser. The Information Retrieval Series, 2006, vol. 20, pp. 1–10.

[86] K. Moilanen and S. Pulman, "Sentiment Composition," in *Proceedings of Recent Advances in Natural Language Processing 2007 (RANLP 2007)*, 2007, pp. 378–382.

[87] A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak, "Determining Negation Scope and Strength in Sentiment Analysis," in *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011)*. IEEE, 2011, pp. 2589–2594.

[88] L. Jia, C. Yu, and W. Meng, "The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM, 2009, pp. 1827–1830.

[89] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment Strength Detection for the Social Web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163–173, 2012.

[90] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghasem-Aghaee, "Sentiment Prediction Based on Dempster-Shafer Theory of Evidence," *Mathematical Problems in Engineering*, 2014. [Online]. Available: http://dx.doi.org/10.1155/2014/361201

[91] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976, vol. 1.

[92] Lun-Wei Ku and Yu-Ting Liang and Hsin-Hsi Chen, "Opinion Extraction, Summarization and Tracking in News and Blog Corpora," in *Proceedings of the AAAI Spring Symposium 2006: Computational Approaches to Analyzing Weblogs (AAAI-SS 2006)*. AAAI, 2006, pp. 100–107.

[93] G. Carenini, R. T. Ng, and A. Pauls, "Multi-Document Summarization of Evaluative Text," in *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL 2006)*. ACL, 2006, pp. 305–312.

[94] H. D. Kim and C. Zhai, "Generating Comparative Summaries of Contradictory Opinions in Text," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM, 2009, pp. 385–394.

[95] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[96] B. C. Wallace, "Computational Irony: A Survey and New Perspectives," *Artificial Intelligence Review*, pp. 1–17, 2013.

[97] I. Peñalver-Martinez, F. Garcia-Sanchez, R. Valencia-Garcia, M. Ángel Rodríguez-García, V. Moreno, A. Fraga, and J. L. Sánchez-Cervantes, "Feature-Based Opinion Mining through Ontologies," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5995 – 6008, 2014.

Kim Schouten is a PhD candidate at the Erasmus University Rotterdam, focusing on aspect-level sentiment analysis and its application, implicitness of aspects and sentiment, and how to move towards a more semantics-oriented form of sentiment analysis. Other topics of interest include the application of language technology within an economic framework, and language in relation to artificial intelligence.

Flavius Frasincar is an assistant professor in information systems at Erasmus University Rotterdam, the Netherlands. He has published in numerous conferences and journals in the areas of databases, Web information systems, personalization, and the Semantic Web. He is a member of the editorial board of the International Journal of Web Engineering and Technology, and Decision Support Systems.