Chapter 2
Sentiment Analysis

Abstract There can, in principle, be no systematic discovery procedure that would yield a set of grammatical rules. The discovery of the rules of language is like the discovery of the rules of nature, a stumbling and unsystematic affair that calls primarily for insight and imagination (Chomsky cited in Burling, Patterns of language: structure, variation, change, 1992, p. 56). This chapter presents a review of the literature in the research field of sentiment analysis to provide an overview of the sentiment lexicon landscape in which to situate SenticNet. The chapter contrasts the concept-based approach to sentiment analysis to linguistics-based approaches, in order to elicit a better understanding of the potential advantages of the concept-based model over the linguistic-based model. The choice of contrasting these two approaches is motivated by their shared interest in integrating semantics into the solution to the problem of detecting sentiment polarity. The near impossibility of producing a complete set of linguistic rules, as suggested by Chomsky, on which to model the detection of sentiment polarity lays the foundations for building the case for the concept-based approach enabled by SenticNet, and further supports the interest in the research question posed in Sect. 1.3. The chapter consists of five main sections. Section 2.1 provides an overview of typical challenges encountered in sentiment analysis. Section 2.2 discusses the levels at which sentiment analysis can be performed. Section 2.3 discusses supervised and unsupervised learning approaches to sentiment analysis. Section 2.4 discusses linguistics-based sentiment analysis. Section 2.5 discusses lexicon-based sentiment analysis.

Keywords Affective knowledge · Linguistics · Machine learning · Part-of-speech · Semantic orientation · Syntactic patterns · Synsets

The goal of this chapter is to contrast the concept-based approach to sentiment analysis to linguistics-based approaches, in order to elicit a better understanding of the potential advantages of the concept-based model over the linguistic-based model. The choice of contrasting these two approaches is motivated by

R. Biagioni, *The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts*, SpringerBriefs in Cognitive Computation 4,
DOI 10.1007/978-3-319-38971-4_2

their shared interest in integrating semantics into the solution to the problem of detecting sentiment polarity. The near impossibility of producing a complete set of linguistic rules, as suggested by Chomsky, on which to model the detection of sentiment polarity lays the foundations for building the case for the concept-based approach enabled by SenticNet, and further supports the interest in the research question posed in Sect. 1.3. In addition, this chapter aims to provide an overview of typical challenges encountered in sentiment analysis and of the sentiment lexicon landscape in which to situate SenticNet.

This chapter presents a review of the literature in the research field of sentiment analysis. The chapter consists of five main sections. Section 2.1 discusses sentiment analysis challenges. Section 2.2 discusses the levels at which sentiment analysis can be performed. Section 2.3 discusses supervised and unsupervised learning approaches to sentiment analysis. Section 2.4 discusses linguistics-based sentiment analysis. Section 2.5 discusses lexicon-based sentiment analysis.

2.1 Sentiment Analysis Challenges

The review of relevant research literature in the field of study of sentiment analysis highlighted the most commonly encountered challenges as being related to the issues of context sensitivity; negation; sarcasm, irony and idioms; the intensity of sentiment; implicit sentiment; compound sentences; ambiguity and domain sensitivity. Each of these challenges is further outlined below.

Context
The sentiment polarity of a word often depends on its context (Ding et al. 2008). For example the word mind is positive in “to be in the right mind” but negative in “to lose one’s mind”. Other words have a lack of sensitivity to context. For example, a word such as “excellent” is positive in almost all contexts.

Negation
Including detection of negation in automated sentiment analysis is of critical importance because negation terms such as “not” change the polarity of opinion words (Liu 2012). For example, in the sentence “the picture quality is not good”, the polarity of the adjective “good” is inverted by the negation term “not”. Negation-handling is not trivial, as not all appearances of explicit negation terms reverse the polarity of the enclosing sentence e.g. “No wonder this is considered one of the best” (Pang and Lee 2008).

Sarcasm, irony and Idioms
Negation can often be expressed in rather subtle ways through sarcasm and irony, which are quite difficult to detect. Such sentences become particularly challenging for sentiment analysis systems when they include sentiment words as illustrated by the sentences “Some cause happiness wherever they go; others whenever they go”, “What a great car! It stopped working in two days” (Pang and Lee 2008; Liu 2012).

Similarly, while most idioms (e.g. "cost someone an arm and a leg") express strong opinions, they are challenging to deal with because of the absence of sentiment words (Ding et al. 2008).

Intensifiers

Intensifiers are terms that change the degree of sentiment (Kennedy and Inkpen 2006). For example, in the sentence "This movie is very good", the terms "very good" have stronger positive polarity than the single word "good" alone. On the other hand, in the sentence "This movie is barely any good", the diminisher "barely" makes the statement less positive.

Implicit opinions and sentiment

Many sentences imply opinions or sentiment without resorting to the use of sentiment words (Hu and Liu 2004). For example, in the sentence "While light, it will not easily fit in pockets", the size of the camera is the opinion target, but the word "size" is not explicitly mentioned in the sentence.

Compound sentences

Compound sentences may express more than one opinion. For example, the sentence, "The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera", expresses both positive and negative opinions (one may say that it has a mixed opinion). For "picture quality" and "battery life", the sentence is positive, but for "viewfinder", it is negative. It is also positive for the camera as a whole.

Ambiguity

A significant challenge in natural language processing is how to handle the ambiguity that arises when interpreting a single sentence. Lexical ambiguity is where different possible meanings are associated with one word, depending on its part-of-speech class. A part-of-speech class is a lexical category such as noun or adjective. For example, the meaning of the word "back" changes depending on whether it is an adverb (e.g. "go back"), an adjective (e.g. "back door"), a noun (e.g. "the back of the room") or a verb (e.g. "back up your files"). Syntactic ambiguity refers to a situation where a sentence may be interpreted in more than one way due to ambiguous sentence structure. For example, "I saw the man with the binoculars" has two interpretations: one where the prepositional phrase "with the binoculars" relates to the noun "man" and one where it relates to the verb "saw". Syntactic ambiguity leads to semantic ambiguity, because "one interpretation means that the man has binoculars and the other means that I used binoculars" (Russell and Norvig 2009). While humans can comprehend the intended meanings by inferring from context or by drawing on personal knowledge and understanding, computers don't benefit from the subtleties of human experience and learning. Resolving ambiguity is one of the most difficult tasks in natural language processing because computers lack "common sense" (PARC Natural Language Processing 2007).

Domain Sensitivity

A key problem in sentiment analysis is its sensitivity to the domain from which either training data is sourced, or on which a sentiment lexicon is built. This is referred to as the problem of domain dependence. A domain refers to the topic of a given text; e.g. cars or movies. The same word in one domain may have positive polarity, while, in another domain, it has negative polarity. For example "unexpected" has positive polarity in the movie domain (e.g. "unexpected plot") but negative polarity in the automotive domain (e.g. "unexpected steering") (Liu 2012). Machine learning methods are highly sensitive to the domain from which the training data has been extracted, and can not be employed for domains in which no training data exists. A classifier trained using opinionated texts from one domain often performs poorly when it is applied to opinionated texts from another domain. With regard to lexical-based methods it is a challenging task to compile general-purpose sentiment lexicons that are capable of performing well for every topic.

2.2 Levels of Analysis

Sentiment analysis can be performed at different levels: document-level, sentence-level, entity-level and aspect-level. Document-level sentiment analysis classifies a whole opinionated document as either positive or negative (Pang et al. 2002; Turney 2002; Godbole et al. 2007; Devitt and Ahmad 2007). Liu (2012) notes that this level of analysis assumes that each document expresses opinions on a single entity (e.g. a single product), and thus is not applicable to documents which evaluate or compare multiple entities. Sentence-level sentiment analysis classifies individual sentences as positive or negative (Bethard et al. 2004; Kim and Hovy 2004). Liu (2010) notes that an assumption of sentence-level sentiment classification is that sentences express a single opinion from a single opinion-holder; e.g. "The picture quality of this camera is amazing". However, as noted in the previous section, compound sentences may express more than one opinion. Entity-level and aspect-level sentiment analysis is based on the idea that sentiments have a target. For example, in the sentence "Although the service is not that great, I still love this restaurant", the sentence adopts a positive stance concerning the entity "restaurant" but a negative stance regarding the aspect of "service". Thus, the goal of this level of analysis is to discover sentiments on entities and/or their features (Yi et al. 2003; Sebastiani et al. 2006).

2.3 Supervised Versus Unsupervised Sentiment Analysis

Sentiment analysis can be performed in a supervised or unsupervised manner. The term "supervised" refers to a methodology whereby applications learn to make classification decisions from labelled examples in training data. Labelled training data refers to data that has been manually annotated, that is, whether a text or sentence expresses positive or negative sentiment. Such applications require large amounts of domain-specific training data in order to learn effectively. Creating labelled training data for all possible domains is expensive and time-consuming. Supervised sentiment analysis is therefore limited by the availability of domain-specific labelled data. The term “unsupervised” refers to a methodology whereby applications classify data without using training data or creating models. Unsupervised sentiment analysis, conversely, does not require labelled training data, and is domain-independent. Typically, unsupervised learning occurs through the clustering, or grouping, of documents or sentences belonging to the same positive or negative sentiment polarity category.

Supervised sentiment analysis approaches are based on finding a suitable set of classification features. Classification features are distinctive attributes in a given text that are suitable for effectively discriminating between positive and negative sentiment. Such features were introduced in Sect. 1.2 as “sources of information” on which sentiment analysis strategies can be based. Taboada et al. (2011) found that supervised methods achieve high accuracy in detecting sentiment polarity in natural language text. Supervised sentiment analysis techniques commonly use the machine learning algorithms Naive Bayes classifier, Maximum Entropy and Support Vector Machine (SVM). These algorithms are described in detail in Alpaydin (2010). Unsupervised sentiment analysis approaches are based on determining the semantic orientation of specific words or phrases within a sentence or document to infer their sentiment orientation. A phrase is a small group of words standing together as a conceptual unit; e.g. “to improve standards”. The two main unsupervised approaches to sentiment analysis are linguistics-based (e.g. Turney 2002) or lexicon-based (e.g. Taboada et al. 2011; Feldman 2013).

2.4 Linguistics-Based Sentiment Analysis

To computers text is a simple sequence of character strings. This makes it challenging to automatically extract the most useful elements contained in text. Linguistics-based sentiment analysis divides text and sentences not only into their constituent words but also identifies their syntactic structure.

Sentiment analysis based on linguistic techniques aims to find part-of-speech patterns that are most likely to express an opinion. Such patterns can take the form of word sequences such as an adjective followed by a noun, or the form of more complex syntactic categories such as noun or verb phrases. These, in turn, can be combined into “trees” representing the nested phrase structure of sentences (Russell and Norvig 2009).

Liu (2012) notes that adjectives have been shown to be important indicators of opinions. This is supported by the observation that several researchers have used adjectives as their primary sentiment detection feature. A brief description of past sentiment analysis research based on adjectives and/or part-of-speech patterns is given below:

- Hatzivassiloglou and McKeown (1997) use a list of adjectives, manually labelled as positive or negative, to correlate the conjunctions "but" or "and" with semantic orientation. Conjunctions are linguistic features that connect words, sentences, phrases or clauses. Hatzivassiloglou and McKeown demonstrated that the conjunctions provide indirect information about sentiment orientation of conjoined words i.e. that they have the same sentiment polarity. For example, the adjectives "fair and legitimate" and "corrupt and brutal" are in each case of identical polarity.
- Turney (2002) classifies sentiment of customer reviews sampled from different domains (automobiles, banks, movies, and travel destinations) using rules based on heuristic part-of-speech patterns that are considered likely to be used to express opinions. The patterns consist of adjectives (JJ), adverbs (RB, RBR, RBS), verbs (VB, VBD, VBN, VBG) and nouns (NN, NNS). This approach is based on the idea that although adjectives often indicate subjectivity, they provide insufficient context to reliably determine sentiment polarity as illustrated in Sect. 1.3.
- Benamara et al. (2007) perform part-of-speech-based sentiment analysis using adverbs and adverb-adjective combinations e.g. "very bad". This approach is based on the observation that adverbs usually act as modifiers of adjectives. A modifier is defined in the Oxford Dictionaries online[1] as "a word, especially an adjective or noun […] that restricts or adds to the sense of a head noun".
- Ding et al. (2008) use adjective, adverb, verb and noun to determine the orientation of opinions about product features in customer reviews. The inclusion of part-of-speech classes other than adjectives is based on the fact that, while orientations apply to most adjectives, there are those adjectives that have no orientations (e.g., external, digital).

2.5 Lexicon-Based Sentiment Analysis

Lexicons that contain compilations of words annotated with sentiment orientation are referred to as sentiment lexicons (Ohana 2009). The rationale underlying lexicon-based techniques is the assumption that the most important indicators of sentiment in natural language text are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, "good", "wonderful" and "amazing" are positive sentiment words, while "bad", "poor" and "terrible" are negative sentiment words (Liu 2012). Sentiment lexicons relate words to sentiment polarity.

---

[1]http://www.oxforddictionaries.com/definition/english/modifier.

The Sentiment Lexicon Landscape

In this section, an overview of past and current sentiment lexicons is presented:

- Stone and Hunt (1963) developed a system for content analysis named "the General Inquirer". The system enabled the understanding of "the psychological forces" in a document. The General Inquirer system used two dictionaries: a purpose-built psycho-sociological dictionary, and an anthropological dictionary used for studying themes in the folktales of different cultures. The former contains category information such as "Persons", "Physical Objects", "Emotions" or "Culture", while the latter contains categories such as "Dominate", "Follow", "Leader", "Power" or "Agree". When analysing a sentence, GI looked up words in the dictionaries and, if a match was found, tags indicating the word's membership in one or more categories specified in the dictionaries were attached to the sentence.
- Esuli and Sebastiani (2005) used entries in the English language dictionary of synonyms WordNet (Kamps et al. 2004) to create the sentiment lexicon SentiWordNet. Starting from the lists of positive and negative words used by Turney and Littman (2003), SentiWordNet was "grown" by repeatedly searching for and adding word synonyms from WordNet. Synonyms in WordNet are grouped into "synsets". Each synset is annotated with a definition, example uses and their relationship with other synsets. Wordnet does not contain sentiment-related information. Terms in SentiWordNet may have multiple "senses" i.e. meanings. For example, the term "cold" can mean "having a low temperature" or "without human warmth or emotion". To allow differentiation between the senses of terms, SentiWordNet provides descriptions, named glosses, for each entry. The fact that terms in SentiWordNet may be stored multiple times, based on their sense, has implications for sentiment analysis applications. For example, the polarity value of the adjective "cold", with the meaning "having a low temperature" (e.g. "cold beer"), is different from the polarity value of the adjective "cold" with the meaning "being emotionless" (e.g. "cold person").
- Strapparava and Valitutti (2004) created the sentiment lexicon WordNet-Affect based on a subset of WordNet synsets. The lexicon consists of "affective knowledge". Affective knowledge is an umbrella term used for words that express moods, feelings and attitudes. In WordNet-Affect, synsets are associated with an affective label. An affective label represents an affective category such as an emotion, feeling, etc. WordNet-Affect is, in effect, a variant of WordNet where synsets have been enriched with affective meaning. For example, in WordNet-Affect, the synset "anger" is attached to the label "emotion".

SentiWordNet and WordNet-Affect are the most widely used sentiment lexicons (Poria et al. 2013). Both lexicons are mostly limited to single-word concepts (Poria et al. 2013).

Limitations of Sentiment Lexicons

Sentiment words provide important clues for sentiment analysis. However, the sole reliance on sentiment lexicons is, as described in Liu (2012) and Pang and Lee (2008), subject to potential limitations:

- Sentiment lexicons are domain-independent. This affects the ability of sentiment analysis systems that rely on such lexicons to deal with the fact that the same term may have opposing polarity, depending on the domain in which it is used. For example, "go read the book" most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews. Dealing with such situations is referred to as word sense disambiguation.
- Domain independence makes sentiment lexicons portable. Portability refers to the ability to apply sentiment lexicons on text from any domain, be it movies, digital cameras or emotion-related texts. However, the downside to portability is that it makes sentiment lexicons less precise when confronted with the disambiguation scenarios described above.
- Similarly to the above issue, words can have different polarities within the same domain. For example, in the camera domain, the word "long" clearly expresses opposite opinions in the following two sentences: "The battery life is long" (positive) and "It takes a long time to focus" (negative).
- Words found in sentiment lexicons do not always express sentiment. For example, in "I am looking for a good car to buy", "good" does not express a positive or negative sentiment.
- Traditional sentiment lexicons cannot deal with implicit (or implied) sentiment. For example, in the sentence "the food is too expensive", the concept of "price" is an implicit information source, as it does not appear in the sentence but it is implied.
- It is difficult to create unique sentiment lexicons covering all possible language constructs. For example, slang is common in online social networks but is not typically supported in sentiment lexicons.

2.6 Conclusion

The sentiment analysis literature review performed in this chapter is relevant to the research question in all aspects. The review of challenges associated with sentiment analysis highlighted the obstacles that are likely going to be faced during the implementation phase of the research and provide valuable insights into the diversity of issues that must be considered when performing sentiment analysis. The review of the levels of analysis of sentiment analysis indicated that testing whether multi-word concepts outperform single-word concepts can be done at either document-, or sentence-level under the condition that the whole document or sentence expresses opinions on a single opinion target. The sentiment analysis methodologies reviewed in Sect. 2.3 indicated that the supervised sentiment analysis approach can be employed under the condition that a suitable pre-labelled dataset is available. Furthermore, the selection of classification features as applied in the supervised approaches can be applied to both concept types. However, the unsupervised lexicon-based approach appeared as the most appropriate approach in light of the fact that the research is specifically focused towards the SenticNet sentiment lexicon.

The review of linguistics-based sentiment analysis complements the position expressed by Chomsky, that it is not possible to harness natural language grammar in an automated way. The research examples described in this chapter illustrate that linguistic rules for identifying sentiment are hand-crafted and heuristic. No universal and exhaustive sentiment grammar exists. This finding provides additional support for lexicon-based approaches in general, and concept-based approaches in particular. While sentiment lexicons are also not exhaustive, they are not restricted by limitations that are out of their control such as the complexities of a language. A lexicon can be as exhaustive as is wanted by its authors. Similarly concepts are free from restrictions imposed by part-of-speech affiliation. In the research examples described in this chapter, approaches relied on adverbs, adjectives, nouns, etc. and sequences thereof. However, combinations of nouns such as "birthday gift" were not attempted because the pattern in itself does not indicate the presence of possible sentiment information. Lastly, the review of the sentiment lexicon landscape illustrates the novel aspects of SenticNet by highlighting the lexical, word-based, nature of traditional sentiment lexicons. SenticNet, in contrast, is a concept-based lexicon in which lexical affinity does not play a role.

After having ascertained aspects of the research question from a field of study point of view, the next step is to investigate and analyse SenticNet.

References

Alpaydin E (2010) Introduction to machine learning, 2nd edn. MIT Press, Cambridge, Mass

Benamara F, Cesarano C, Picariello A, Reforgiato D, Subrahmanian VS (2007) Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: Proceedings of International Conference Weblogs Social Media ICWSM

Bethard S, Yu H, Thornton A, Hatzivassiloglou V, Jurafsky D (2004) Automatic extraction of opinion propositions and their holders. In: In 2004 AAAI spring symposium on exploring attitude and affect in text, pp. 22–24

Burling R (1992) Patterns of Language: structure, variation, change, 1st edn. Academic Press Inc., San Diego, California

Devitt A, Ahmad K (2007) Sentiment polarity identification in financial news: a cohesion-based approach. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics

Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 international conference on web search and data mining. ACM, pp. 231–240

Esuli A, Sebastiani F (2005) Determining the semantic orientation of terms through gloss classification. In: Proceedings of the 14th ACM international conference on information and knowledge management, CIKM'05. ACM, New York, NY, USA, pp 617–624. doi:10.1145/1099554.1099713

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56:82. doi:10.1145/2436256.2436274

Godbole N, Srinivasaiah M, Skiena S (2007) Large-scale sentiment analysis for news and blogs. In: Proceedings of the international conference on weblogs and social media (ICWSM)
Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics, ACL'98. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 174–181. doi:10.3115/976909.979640
Hu M, Liu B (2004) Mining opinion features in customer reviews. In: AAAI, pp 755–760
Kamps J, Marx MJ, Mokken RJ, De Rijke M (2004) Using wordnet to measure semantic orientations of adjectives
Kennedy A, Inkpen D (2006) Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell 22:110–125
Kim S-O, Hovy E (2004) Determining the sentiment of opinions. Presented at the proceedings of the COLING conference, Geneva
Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of natural language processing, 2nd edn. Taylor and Francis Group, Boca
Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5:1–167. doi:10.2200/S00416ED1V01Y201204HLT016
Ohana B (2009) Opinion mining with the SentWordNet lexical resource. Dublin Institute of Technology
Pang B, Lee L (2008) Opinion mining and sentiment analysis
Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing-volume 10. Association for Computational Linguistics, pp 79–86
PARC Natural Language Processing (WWW Document) (2007) http://www.parc.com/content/attachments/naturallanguage_backgrounder_parc.pdf. Accessed 16 Jan 2015
Poria S, Gelbukh A, Hussain A, Howard N, Das D, Bandyopadhyay S (2013) Enhanced senticnet with affective labels for concept-based opinion mining. IEEE Intell Syst 28:31–38. doi:10.1109/MIS.2013.4
Russell S, Norvig P (2009) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall, Upper Saddle River
Sebastiani AF, Esuli A, Sebastiani F (2006) Determining term subjectivity and term orientation for opinion mining andrea Esuli. In: In Proceedings of the 11th conference of the european chapter of the association for computational linguistics (EACL'06
Stone PJ, Hunt EB (1963) A computer approach to content analysis: studies using the general inquirer system. In: Proceedings of the May 21–23, 1963, spring joint computer conference, AFIPS'63 (Spring). ACM, New York, NY, USA, pp 241–256. doi:10.1145/1461551.1461583
Strapparava C, Valitutti A (2004) WordNet affect: an affective extension of wordnet. In: LREC, pp 1083–1086
Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput. Linguist. 37:267–307
Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 417–424
Turney PD, Littman ML (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Trans Inf Syst 21:315–346. doi:10.1145/944012.944013
Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: In IEEE International Conference on Data Mining (ICDM, pp 427–434)

Springer

http://www.springer.com/978-3-319-38970-7

The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts
Biagioni, R.
2016, VI, 55 p. 13 illus., 8 illus. in color., Softcover
ISBN: 978-3-319-38970-7