International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 7 Issue 5, May-2017, pp. 99~104
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196 | Thomson Reuters Researcher ID: L-5236-2015

Foundations and trends in information retrieval using Opinion Mining and Sentiment Analysis

Shweta Rana
Assistant Professor
Amity University Haryana

Abstract
Opinion mining is a relatively new and challenging field dedicated to detecting subjective content in text documents, with a variety of uses in real world applications It is a new and exciting field of research concerned with extracting opinion related information from textual data sources. It has the potential for a number of interesting applications both in commerce and academic areas, and poses novel intellectual challenges, which continues to attract considerable research interest. In this paper the opinion mining is introduced, its motivations, key tasks and challenges are discussed in more details. Then, the SentiWordNet lexical resource for opinion mining is presented, and its potential advantages, applications and limitations are discussed.

I. Opinions in Text

Information concerning people's opinions can be a very important component for more accurate decision making in a number of domains. Companies, for instance, have a keen interest in finding out what are their customers' opinions on a new product launched on a marketing campaign. Consumers on the other hand would benefit from accessing other people's

opinions and reviews on a given product they are intending to purchase, as recommendations from other users tend to play a part on influencing purchasing decisions. Knowledge of other people's opinions is also important in the political realm, where for instance, one could find out the sentiment towards a new piece of legislation, or an individual such as a politician or activist. In recent years, the
internet has enabled access to opinions in the form of written text from a variety of sources and in a much larger scale. It also made it easier for people to express their opinions on virtually any subject by means of specialized product review websites, discussion forums and blogs. It is worth highlighting that consumer goods are not the only target of opinion related content: specialized websites that gather and provide opinion information on companies, politicians and education resources are also available. Opinion sources are not restricted to specialized review sites, and are also contained in users' blog posts, discussion forums and embedded in online social networks. The internet is clearly a vast repository of publicly available user generated content dedicated to expressing opinions on any topic of interest. However, despite the clear benefit of having such information available, it was also pointed out that 58% of internet users reported finding product information online either confusing, difficult to find, or were overwhelmed by the volume of information available. Automated methods for efficiently extracting knowledge from these resources appear an attractive proposition for both individuals who would be able to make better decisions and to companies who could quickly gauge. In addition, opinions are generally expressed in textual form, making it a rich ground for the application of text mining and techniques to analyze natural language. Thus, the motivating need to analyze large volumes of opinion information, coupled with advances in natural language processing and machine learning methods gave rise to research in the emerging field of Opinion Mining. Opinion Mining is concerned with applying computational methods for the detection and measurement of opinion, sentiment and subjectivity in text. A text document can be seen as a collection of objective and subjective statements, where objective statements refer to factual information present in text, and subjectivity relates to the expression of opinions, evaluations and speculations. To further illustrate the motivations for performing

opinion mining, we now survey the potential applications of computing systems that apply techniques that detect and extract the subjective aspects of text.

Search Engines

The most direct application of opinion mining techniques would be the searching of opinions within documents. Finding out subjective statements related to a topic, and their bias can augment traditional search engines into recommendation engines by retrieving results on a given topic containing only positive or negative sentiment, for example when searching for products that received good reviews on a particular area, like a user query for digital cameras with good feedback on battery life.

Inappropriate Content

In a collaborative environment such as a discussion group or email list, opinion mining could be applied to classify subjective statements containing overly heated or inappropriate remarks, also called flaming behavior. Similar techniques could assist more efficient online advertisement strategies by avoiding ad placements next to content that is related to the ad campaign, but carries unfavorable opinions towards a certain product or brand.

Customer Relationship Management

Systems that manage customer interactions can become more responsive by using sentiment detection as a tool to automatically predict the level of satisfaction of client feedback. One example is the automatic classification of customer feedback replies by email containing positive or negative sentiment, which could then be used for automatically routing of messages into the appropriate teams for corrective actions when necessary.

Business Intelligence

Opinion mining has the ability to add analysis of subjective components of text to discover new knowledge from data. This may take the form of aggregated sentiment bias information from user feedback which can be used to drive marketing campaigns and improve product design. In the financial industry, sentiment information present on financial news have been studied to assess its impact on the performance of securities.

Advantages of Knowledge Management

From the examples of opinion mining applications presented above, it can be seen this field of research has the potential to add value to knowledge management efforts in companies across a range of knowledge based activities. Knowledge based systems that store explicit content can become more efficient by extending its query interface to include opinion information for more relevant results, or excluding subjective documents when more factual results are needed. Knowledge sharing systems that provision collaborative environments for exchange of explicit or tacit knowledge can become more fluid and require less administration efforts by employing sentiment detection to avoid flaming and other unwanted user behavior; finally, knowledge discovery systems can leverage opinion information to help knowledge creation in the organization, and improve decision making where user feedback is relevant.

II. Key Problems Addressed by Opinion Mining

In an attempt to map the activities of the emerging field of opinion mining the research survey categorizes the area into two broad fields of classification and extraction. Classification would entail research related to detecting in first instance if a piece of text can be categorized as subjective or objective and, in case it is subjective, be ability to correctly predict the text's sentiment orientation or polarity; the extraction aspect of opinion mining shares the concerns of information retrieval, and attempts to identify within a text document what are the key attributes of an opinion, such as the holder or to what entity it refers to, with a view to build summarizes based on opinion information. Agreement detection is a similar problem where differing or agreeing opinions between two distinct documents are sought. One field of research that shares some of the concerns of opinion mining is that of affective computing, aiming at the development of computational approaches for detecting human emotions such as anger, fear and humor. Affective computing has applications in human computer interaction, but is closely linked to the problem of detecting subjective text, since both relate to the expression of human emotions. For the purposes of this research, the focus or this review is on the predictive aspects of opinion mining related to the tasks of subjectivity detection and sentiment

classification of texts. The subjectivity of a sentence is defined based upon previous work in linguistics and literary theory. First, there are subjective elements: the linguistic expressions that characterize private states of mind. Characterizing subjective elements is not a trivial task; they may appear in text as single words, expressions or entire sentences, may depend on the context, and may also be evident in text style. A subjective element expresses the opinions, thoughts and speculations of a source, that is the document author or someone mentioned in the text. Finally, a subjective element has a target, or the object being referred to. Subjectivity detection is generally concerned with finding the subjective elements or sentiment in text, but other aspects of the above characterization can also be of relevance for query systems and summarization tasks. For instance, when tracking the opinions of a given person. Subjectivity is annotated manually at expression, sentence and document levels, and used to train detection methods based on terms presence and term collocations, or their position in the text relative to each other. This is based on the hypothesis that subjectivity of an expression is a function of how subjective their surrounding elements in text are.

Sentiment Classification

Sentiment classification is concerned with determining what, if any, is the sentiment orientation of the opinions contained within a given document. It is assumed in general that the document being inspected is known to represent opinion, such as a product review. Opinion orientation can be classified as belonging to opposing positive or negative polarities – positive or negative feedback about a product, favorable or unfavorable opinions on a topic – or ranked according to a spectrum of possible opinions, as is the case with film reviews with feedback ranging from zero to five stars.

Word Vectors

One natural approach to performing sentiment classification is to take the traditional text mining representation of documents as word vectors, where each entry maps to a term found in the corpus of documents, and the value of a given entry corresponds to a measure of term presence or a measure of relative term frequency from the field of text mining and information retrieval. In a series of experiments using various classes of word vectors for sentiment classification of film reviews generated positive results for single term word vectors – or unigrams - using binary presence values for each term. Binary presence did perform better than frequency-based word vectors, suggesting that term existence, rather than frequency is more significant to opinion identification.

Another similar experiment based on word vectors and product reviews as the data set reports good results for tri-grams is seen. Taking the traditional text mining approach to train a classifier based on word vectors for opinion mining generates good classification performance results, but these results stay well below those obtained for topic-based document classification using the same techniques. Empirical performance metrics for topic-based text categorization using Support Vector Machines show how high precision, high recall topic based classification can be achieved, based on results using well known experiment data sets. This observation, coupled with further analysis of opinion bearing documents suggests that sentiment information needs to be captured by other means.

“This film should be brilliant. It sounds like a great plot, the actors are first grad. However, it can’t hold up.”

In the above case a sentence contains a high number of positive statements, building up the expectation of a positive review, but the overall sentiment of the review is still negative. This affects prediction decisions based on term information presence alone, and suggests that the order of which opinions are presented is of importance to overall sentiment.

Word Sense Disambiguation

Subjectivity detection is improved by adding a subjective feature to detect terms in need of disambiguation, and the authors speculate improvements to sentiment classification tasks with the assistance of term disambiguation techniques. This need has also been highlighted when inspecting results of sentiment classification experiment based on term opinion information.

Parts of Speech

Classifying terms from a textual document into its grammatical roles, or parts of speech within a sentence has also been explored in opinion mining. A motivating factor behind this approach is that

detecting parts of speech can be considered a form of word disambiguation for the cases where word senses are associated with its grammatical use, such as noun, verb, etc. Another factor is the finding that adjectives are considered good indicators of opinion information and have been seen to provide good correlation to sentiment orientation. In a study reports good results using only adjective words as features to perform sentiment classification using a machine learning method, however with poorer results than using full word vectors as features. The use of parts of speech as a pre-processing step for deriving features for opinion mining has also been seen in a number of other sentiment classification experiments.

Combining Approaches

Taking the view that different methods for performing sentiment classification capture different types of sentiment related information from a document, it is worth noting the contribution in the literature to combining results from more than one classifier in order to obtain better results. This can be done not only to address induction bias from a specific classifier algorithm, but also to make better decisions from a pool of classification techniques, each leveraging different types of data. A similar approach is seen with the combination of proximity metrics and term relationships extracted from a lexicon.

One common approach in performing both subjectivity detection and sentiment classification involved the use of key words that are assumed to be indicative of either positive or negative bias, and therefore also of overall subjectivity. This idea is based on the hypothesis that words can be considered as a unit of opinion information, and several methods based on this assumption have been proposed with considerable success. One interesting aspect of approaches based on word lists is that it does not necessarily require training data for making predictions, since it relies only on a pre-defined sentiment lexicon, thus being applicable to cases where no training data is present. For this reason, these methods are often labeled as unsupervised learning approaches.

Creating word lists manually however time consuming is, and approaches have been proposed in the literature for automatically creating resources that contain opinion information on words based on readily available lexicons.

Senti WordNet

One example of a lexical resource conceived to assist in opinion mining tasks is SentiWordNet. SentiWordNet aims at providing term level information on opinion polarity by deriving this information from the WordNet database of English terms and relations in a semi-automatic fashion. For each term in WordNet, a positive and a negative score ranging from 0 to 1 is present in SentiWordNet, indicating its polarity, with higher scores indicating terms that carry heavy opinion bias information, whereas lower scores indicate a term being less subjective.

WordNet

WordNet is a lexical database for the English language where terms are organized according to their semantic relations. It has been widely applied to problems in natural language processing. Before describing how SentiWordNet is built, a brief discussion on the database that originated it will be of help in understanding the underlying motivations and how the data is organised. The WordNet lexicon is the result of research efforts in linguistics and psychology at Princeton University on better understanding the nature of semantic relations of terms in the English language, and on providing a complete lexicon in the English language where terms can be retrieved and explored according to concepts and their semantic relationships. At its third version, WordNet is available as a database, searchable via web interface or via a variety of software APIs, providing a comprehensive database of over 150.000 unique terms organized into more than 117,000 different meanings. WordNet also grew with extensions of its structure applied to a number of other languages.

Key Term Relationships

The key relation between terms in WordNet is similarity of meaning, or synonymy. Terms are grouped together into sets of synonyms called synsets. The general criteria for grouping terms together into a synset is whether a term used within a sentence on a specific context can be replaced by another term on the same synset without modifying the sentence's understanding. One direct implication of this structure is that terms must also be differentiated by syntactic categories, since nouns, adjectives verbs and adverbs are not interchangeable within a sentence. Synsets also

contain a short descriptive text defining its terms – or gloss – to assist in specifying its meaning. This is particularly useful on synsets with only a single term, or synsets with a small number of relations. Another important term relationship present in WordNet is anonymity, or whether terms are conceptually opposites. In the special case of adjectives, there is a distinction between direct and indirect antonyms, or when terms can be categorized as direct opposites, or indirectly via another conceptual relationship. The words "wet/dry" are qualified as direct antonyms, however "heavy/weightless" are conceptually opposites and thus indirect antonyms, since they belong to synsets where a direct antonym exists between the terms ("heavy/light") but are not directly correlated. Hyponymy is another class of relationship present in WordNet, and indicates a hierarchical "is-a" type of relationship between terms, such is the case with "oak/plant" and "car/vehicle", while meronymy relationships indicate "part-of" types of relationship between terms. For the special case of adjectives, an attribute type of relationship exists, indicating of what generic attribute the adjective is a modifier, for example the example adjectives "heavy" and "light" are modifiers of the attribute "weight". WordNet would then link the noun representing the attribute to the adjectives that modify it with this type of relationship.

Building SentiWordNet

Building on the strengths of WordNet's semantic relationships, SentiWordNet derives opinion scores for synsets using a semi-supervised method where only a small portion of synset terms - called the paradigmatic terms - are manually labeled, with the remaining database derived using an automated method. The complete process is described in and summarized below:

1. Manually label paradigmatic terms extracted from the WordNet-Affect lexical resource into positive or negative labels, according to opinion polarity.
2. Iteratively expand each label by adding terms from WordNet that are connected to already labeled terms by a relationship considered to reliably preserve term orientation. The following relationships are used to extend the labels:

a. Direct antonym
b. Attribute
c. Hyponymy (pertains-to and derive-from)
d. Also-see
e. Similarity

3. From newly added terms, add to opposite label the terms containing directly opposite opinion orientation, according to the direct antonym relationship.
4. Repeat steps 2 and 3 for a fixed number of iterations K.

Upon completion of steps 1-4, a subset of WordNet synsets is now labeled either positive or negative. To complete the score assignment for all terms, a set of classifiers is trained on their synset glosses, or textual definitions of each synset meaning available on WordNet. The process continues by classifying new entries according to this training data, and generating an aggregated score, as detailed below:

5. For each labeled synset from steps 1-4, produce a word vector representation, along with a positive/negative label. This data set is used to train a committee of classifiers built as follows:

a. Train a pair of classifiers to make the following predictions:

Positive/non positive, and negative/non-negative. Synsets that belong to both positive and negative labels are excluded from the training set and assigned to the "objective" class, with zero-valued positive and negative scores.

b. Repeat process for different sizes of training sets. These are obtained by varying K in the previous stage: 0,2,4 and 6.

c. For each training set, use Rocchio and Support Vector Machine classification algorithms.

6. When applying the set of classifiers to new terms, each resulting classifier returns a prediction score as a result. These summed together and normalized to 1.0 to produce the final positive and negative scores for a term.

The process for building SentiWordNet illustrated above highlights the reliance of term scores on two distinct factors: the choice of paradigmatic words that will generate the initial set of positive and negative scores must be carefully considered, since the extension of scores to the remainder of WordNet terms relies on this core set of terms for making a scoring decision. Secondly, the process relies on synset's textual description, or glosses, for the machine learning stage of the process, to derive a new term's similarity to positive or negative terms.

Applying SentiWordNet

Earlier in this section the advantages of lexicon-based approaches to opinion mining were observed, and results from experiments on both subjectivity detection and sentiment classification were investigated. The use of SentiWordNet as a lexical resource for opinion mining could be of advantage on various instances. The approach of using individual terms as a unit for sentiment information has received considerable research attention in opinion mining, and SentiWordNet could be applied as a replacement to manually building sentiment lexicons from WordNet, often done on an ad-hoc basis for specific opinion mining research, as found on. Validating automated methods for building term orientation information such as SentiWordNet can be useful in the scalability and automation of these approaches to opinion mining.

Conclusion

In this paper opinion mining was surveyed. Opinion mining is a new field of research leveraging components from data mining, text mining and natural language processing, and a wide range of applications of extracting opinion from documents is possible. These range from improving business intelligence in organizations to information retrieval systems, recommender systems
and more efficient online advertising and spam detection. It was seen that opinion mining can be beneficial to knowledge management initiatives either directly, by improving the quality of knowledge repositories through opinion-aware features, or by adding to the knowledge that can be extracted from textual data sources, thus indirectly creating more opportunities for knowledge creation within the company. Finally, the WordNet and SentiWordNet lexical resources were introduced, with a presentation of its building blocks and potential uses. SentiWordNet is an extension of the popular WordNet database of terms and relationships, and is a readily available lexical resource of term sentiment information, which could be used on opinion mining research where a number of similar approaches were devised in an ad-hoc fashion .

References

[1] Boiy E, Hens P, Deschacht K, Moens M, (2007) "Automatic Sentiment Analysis in On-line Text"/

{2} Corney M. de Vel, O., Anderson A, Mohay G. (2002) "Gender-preferential text mining of e-mail discourse".

[3] Csomai A, Rosenzweig J, Mihalcea R. (2007) “WordNet Bibliography”.

[4] Dave K, Lawrence S, Pennock D. (2003) “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification in Product Reviews”.

[5] Cui H, Mittal V, Datar M. (2006) “Comparative Experiments on Sentiment Classification for Online Product Reviews”/

[6] Hotho A, Nurnberger A, Paaß G. (2005) “A Brief Survey of Text Mining”.

[7]Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.

[8] Kamal Nigam, Matthew Hurst. 2004. “Towards a Robust Metric of Opinion.”.

[9]Jeonghee Yi, Tetsuya Nasukawa. 2004. "Sentiment Analysis: Capturing Favorability Using