Statistical Approaches to Concept-Level Sentiment Analysis

Erik Cambria, *National University of Singapore*

Björn Schuller, *Technische Universität München*

Bing Liu, *University of Illinois at Chicago*

Haixun Wang, *Microsoft Research Asia*

Catherine Havasi, *Massachusetts Institute of Technology*

Between the dawn of civilization through 2003, there were just a few dozen exabytes of information on the Web. Today, that much information is created weekly. The advent of the Social Web, in fact, has provided people with new tools—such as forums, blogs, social networks, and content-sharing services—which allow them to create and share in a time- and cost-efficient way their own content, ideas, and opinions with virtually millions of people connected to the World Wide Web. This huge amount of useful information, however, is mainly unstructured (because it's specifically produced for human consumption) and, hence, it isn't directly machine-processable. The opportunity to capture the opinions of the general public about social events, political movements, company strategies, marketing campaigns, and product preferences has raised more and more interest both in the scientific community, for the exciting open challenges, and in the business world, for the remarkable fallouts in marketing and financial market prediction.

Existing Approaches

Existing approaches to sentiment analysis can be grouped into three main categories: keyword spotting, lexical affinity, and statistical methods. Keyword spotting is the most naive approach and probably also the most popular because of its accessibility and economy. Text is classified into affect categories based on the presence of fairly unambiguous affect words like *happy*, *sad*, *afraid*, and *bored*. The weaknesses of this approach lie in two areas: poor recognition of affect when negation is involved and reliance on surface features.

As an example of the first weakness, while the approach can correctly classify the sentence "today was a happy day" as being happy, it's likely to fail on a sentence like "today wasn't a happy day at all." Regarding second weakness, the approach relies on the presence of obvious affect words that are only surface features of the prose. In practice, a lot of sentences convey affect through underlying meaning rather than affect adjectives. For example, the text "My husband just filed for divorce and he wants to take custody of my children away from me" certainly evokes strong emotions, but uses no affect keywords, and therefore, cannot be classified using a keyword-spotting approach.

Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious affect words, it assigns arbitrary words a probabilistic *affinity* for a particular emotion. For example, "accident" might be assigned a 75 percent probability of indicating a negative affect, as in "car accident" or "hurt by accident." These probabilities are usually trained from linguistic corpora. Though often outperforming pure keyword spotting, there are two main problems with the approach. First, lexical affinity, operating solely on the word-level, can easily be tricked by sentences like "I avoided an accident" (negation) and "I met my girlfriend by accident" (other word senses). Second, lexical affinity probabilities are often biased toward text of a particular genre, dictated by the source of the linguistic corpora. This makes it difficult to develop a reusable, domain-independent model.

Statistical methods, such as Bayesian inference and support vector machines, have been popular for affect classification of texts. By feeding a machine learning algorithm a large training corpus of affectively annotated texts, it's possible for the system to not only learn the affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence of other arbitrary keywords (like lexical affinity), punctuation, and word co-occurrence frequencies. However, traditional statistical methods are generally semantically weak, meaning that, with the exception of obvious affect keywords, other lexical or co-occurrence elements in a statistical model have little predictive value individually. As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input. So, while these methods may be able to affectively classify user's text on the page- or paragraph-level, they do not work well on smaller text units such as sentences or clauses.

> In practice, a lot of sentences convey affect through underlying meaning rather than affect adjectives.

Concept-Level Techniques

In light of these considerations, this special issue focuses on the introduction, presentation, and discussion of concept-level techniques to opinion mining and sentiment analysis that are based on statistical approaches. The main motivation for the special issue is to go beyond a mere word-level analysis of text and provide novel approaches to opinion mining and sentiment analysis that allow a more efficient passage from (unstructured) textual information to (structured) machine-processable data, in potentially any domain. For this special issue, we received 30 articles, of which five were carefully selected.

The first article, "Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification" by Riu Xia and his colleagues—which was handled independently during the review process—opens the special issue with a discussion on the domain adaptation problem, which arises often in the field of sentiment classification. In the problem of domain adaptation, there are two distinct needs: labeling and instance adaptation. However, most current research focuses on labeling adaptation, while it neglects instance adaptation. In this article, a comprehensive approach, named feature ensemble plus sample selection (SS-FE), is proposed. SS-FE takes both types of adaptation into account: a feature ensemble (FE) model is first adopted to learn a new labeling function in a feature reweighting manner, and a principal component analysis sample selection (PCA-SS) method is then used as an aid to FE.

THE AUTHORS

Erik Cambria is a research scientist in the Cognitive Science Programme, Temasek Laboratories, National University of Singapore. His research interests include AI, the Semantic Web, natural language processing, and big social data analysis. Cambria has a PhD in computing science and mathematics from the University of Stirling. He is on the editorial board of Springer's *Cognitive Computation* and is the chair of many international conferences such as Brain-Inspired Cognitive Systems (BICS) and Extreme Learning Machines (ELM). Contact him at cambria@nus.edu.sg.

Björn Schuller leads the Machine Intelligence and Signal Processing group at the Institute for Human-Machine Communication at the Technische Universität München. His research interests include machine learning, affective computing, and automatic speech recognition. Schuller has a PhD in electrical engineering and information technology from the Technische Universität München. Contact him at schuller@tum.de.

Bing Liu is a professor of Computer Science at the University of Illinois at Chicago (UIC). His research interests include opinion mining and sentiment analysis, Web mining, and data mining. Liu has a PhD in artificial intelligence from the University of Edinburgh. He has served as the associate editor of *IEEE Transactions on Knowledge and Data Engineering* (TKDE), the *Journal of Data Mining and Knowledge Discovery* (DMKD), and *KDD Explorations*. Contact him at liub@cs.uic.edu.

Haixun Wang is a researcher at Microsoft Research Asia in Beijing, China. His research interests include data management, graph systems, data mining, semantic networks, and text analytics. Wang has a PhD in computer science from the University of California, Los Angeles. He is the associate editor of *IEEE Transactions of Knowledge and Data Engineering* (TKDE) and the *Journal of Computer Science and Technology* (JCST). Contact him at haixunw@microsoft.com.

Catherine Havasi is a cofounder of the Open Mind Common Sense project at the Massachusetts Institute of Technology (MIT) Media Lab, where she works as a postdoctoral associate. Her research interests include commonsense reasoning, dimensionality reduction, machine learning, language acquisition, cognitive modeling, and intelligent user interfaces. Havasi has a PhD in computer science from Brandeis University. Contact her at havasi@media.mit.edu.

In "Retrieving Product Features and Opinions from Customer Reviews," Lisette García-Moya and her colleagues introduce a new methodology for the retrieval of product features and opinions from a collection of free-text customer reviews about a product or service. Such a methodology relies on a language-modeling framework that can be applied to reviews in any domain and language provided with a seed set of opinion words. The methodology combines both a kernel-based model of opinion words (learned from the seed set of opinion words) and a statistical mapping between words to approximate a model of product features from which the retrieval is carried out.

Then, in "Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling," Giuseppe Di Fabbrizio and his colleagues present Starlet, a novel approach to extractive multidocument summarization for evaluative text that considers aspect rating distributions and language modeling as summarization features. Such features encourage the inclusion of sentences in the summary that preserve the overall opinion distribution expressed across the original reviews and whose language best reflects the language of reviews. The article demonstrates how the proposed method offers improvements over traditional summarization techniques and other approaches to multidocument summarization of evaluative text.

Next, in "Multimodal Sentiment Analysis of Spanish Online Videos," Verónica Pérez Rosas and her colleagues consider multimodal sentiment analysis based on linguistic, audio, and visual features. A database of 105 Spanish videos of 2 to 8 minutes in length, containing 21 male and 84 female speakers, was collected randomly from the social media website YouTube and annotated by two labelers for ternary sentiment. This led to 550 utterances and approximately 10,000 words. The joint use and analysis of linguistic, audio, and visual features leads to a significant improvement over the use of each single modality. This is further confirmed when the approach is applied to a set of English videos.

The last article, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," by Martin Wöllmer and his colleagues (and reviewed independently) also discusses multimodal sentiment analysis in online videos. The authors introduce the Institute for Creative Technologies' Multi-Modal Movie Opinion (ICT-MMMO) database of personal movie reviews collected from YouTube (308 clips) and ExpoTV (78 clips). The final set contains 370 of these 1–3 minute English clips in ternary sentiment annotation by one to two coders. The feature basis is formed by 1,941 audio features, 20 video features, and different textual features for selection. Then, different levels of domain-dependence are considered: in-domain analysis, cross-domain analysis based on the 100,000 textual Metacritic movie review corpus for training, and use of online knowledge sources. This shows that cross-corpus training works sufficiently well, and language-independent audiovisual analysis is competitive with linguistic analysis.

These articles are a solid and varied representation of some of the exciting challenges and solutions

emerging in this field. We hope that you enjoy the special issue and that this research fosters future innovations.

Acknowledgments

We would like to thank the Editor in Chief, Daniel Zeng, for his help with this special issue and the 68 reviewers who not only helped with the decision process, but contributed with excellent reviews to make this issue special: Abhishek Jaiantilal, Alexandra Balahur, Alexey Solovyev, Amac Herdagdelen, Andreas Hotho, Antonio Reyes, Appavu Balamurugan, A.R. Balamurali, Carlo Strapparava, Carmen Banea, Cristina Bosco, Cyril Joder, Daniel Olsher, Danushka Bollegala, Dennis Clark, Efstratios Kontopoulos, Eugene Bann, Felix Burkhardt, Felix Weninger, Fernando Fernandez-Martinez, Florian Eyben, Giovanni Acampora, Girgori Sidorov, Henry Anaya-Sanchez, Hidenao Abe, Huan Huang, Jane Malin, Jorge Carrillo de Albornoz, Jose Antonio Troyano, Jose Barranquero, Juan Augusto, Jun Deng, Kalina Bontcheva, Karthik Dinakar, Ken Arnold, Khiet Truong, Laura Plaza, Laurence Devillers, Leslie Fife, Ling Chen, Maarten van der Heijden, Mandy Dang, Marcelo Armentano, Marchi Erik, Mariel Ale, Matthew Aitkenhead, Mehdi Adda, Mitsuru Ishizuka, Mohd Helmy Abd Wahab, Nikolaos Engonopoulos, Paolo Gastaldo, Paolo Rosso, Rada Mihalcea, Richard Crowder, Rodrigo Agerri, Sameera Abar, Serge Sharoff, Stefan Siersdorfer, Stefan Steidl, Stephen Poteet, Vered Aharonson, Wen Xiong, Wen-Han Chao, Wenjing Han, Yair Neuman, Yongzheng Zhang, Zixing Zhang, and Zornitsa Kozareva.