that, as KLD, are not linear functions of individual instances.

We hope to report the results of experimenting with this approach on sentiment quantification data sets in the near future. Concerning the optimization of *ordinal* quantification, instead, further research is still needed to devise ordinal regression methods that can explicitly optimize EMD.

References

1. T. Macer, M. Pearson, and F. Sebastiani, "Cracking the Code: What Customers Say, in their own Words," *Proc. 50th Ann. Conf. Market Research Soc.* (MRS 07), MRS, 2007.
2. D. Giorgetti and F. Sebastiani, "Automating Survey Coding by Multiclass Text Categorization Techniques," *J. Am. Soc. Information Science and Technology*, vol. 54, no. 14, 2003, pp. 1269–1277.
3. G. Forman, "Quantifying Counts and Costs via Classification," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, 2008, pp. 164–206.
4. Y. Rubner, C. Tomasi, and L.J. Guibas, "A Metric for Distributions with Applications to Image Databases," *Proc. 6th Int'l Conf. Vision* (ICCV 98), IEEE CS Press, 1998, pp. 59–66.
5. T. Joachims, "A Support Vector Method for Multivariate Performance Measures," *Proc. 22nd Int'l Conf. Machine Learning* (ICML 05), ACM Press, 2005, pp. 377–384.

Andrea Esuli is a researcher at ISTI-CNR. He has a PhD in information engineering from the University of Pisa, Italy. Contact him at andrea.esuli@isti.cnr.it.

Fabrizio Sebastiani is a senior researcher at ISTI-CNR. He has a "Laurea" degree in computer science from the University of Pisa, Italy. Contact him at fabrizio.sebastiani@isti.cnr.it.

Intelligent Feature Selection for Opinion Classification

Ahmed Abbasi, *University of Wisconsin-Milwaukee*

Although text opinion mining involves many important tasks, accurately assigning sentiment polarities (such as positive, negative, or neutral) and intensities (such as high or low) remains a critical challenge. Given the complexities and nuances associated with opinion classification, it is generally considered more difficult than traditional text mining tasks such as topic-based document categorization. Consequently, prior sentiment-analysis studies have used more sophisticated feature representations, well beyond bag-of-words and word n-grams. The features used include part-of-speech tag n-grams, syntactic phrase patterns, legomena-based collocations, as well as manually and semiautomatically constructed syntactic and semantic phrase patterns and lexicons.[1,2,5] Although these features represent potentially important sentiment discriminators, incorporating them in unison can produce feature spaces spanning tens of thousands of attributes, a situation resulting in the age-old conundrum of disentangling quality from quantity. In addition to the obvious ramifications pertaining to computational feasibility, we must also consider the trade-offs between representational richness and noise, between generalization ability and over-fitting (memorization). Without appropriate feature-selection mechanisms, using large heterogeneous feature spaces is analogous to "throwing the kitchen sink."[3]

This problem is exacerbated by the lack of feature-selection methods specifically crafted for opinion classification. Most existing feature-selection methods are generic techniques that are uniformly applied to input feature value matrices. Examples include information gain, log likelihood, chi squared, and decision-tree models.[2,3,4,8] When applied to text, these methods are often more artificial than they are intelligent. Text features are multidimensional in terms of their informational composition.[4] In addition to various occurrence measures (such as presence and frequency), they encompass lexicology and morphology-based characteristics (including semantics and syntax). There is a need for intelligent feature-selection (IFS) methods that can exploit the syntactic properties of text features while simultaneously leveraging relevant sentiment-related semantic information.

An excellent example of a feature-selection approach tailored to sentiment analysis that utilizes the syntactic relations between text attributes is feature subsumption hierarchies (FSH).[1] Given a set of word n-grams and syntactic n-gram patterns, FSH uses the idea of performance-based feature subsumption to remove redundant or irrelevant higher order n-grams. For instance, only the word bigrams and trigrams that provide additional information (measured using some heuristic) over the unigrams they encompass are retained.[1] For example, the bigram "I like" may be subsumed by the unigram "like," but "basket case" may be retained because it contains important sentiment information not provided by "basket" or "case" alone.

Inspired by FSH, this article presents an IFS approach that incorporates syntactic and semantic information. The proposed approach helps illustrate how rich, heterogeneous feature sets, coupled with appropriate feature-selection mechanisms, can improve opinion-classification performance.

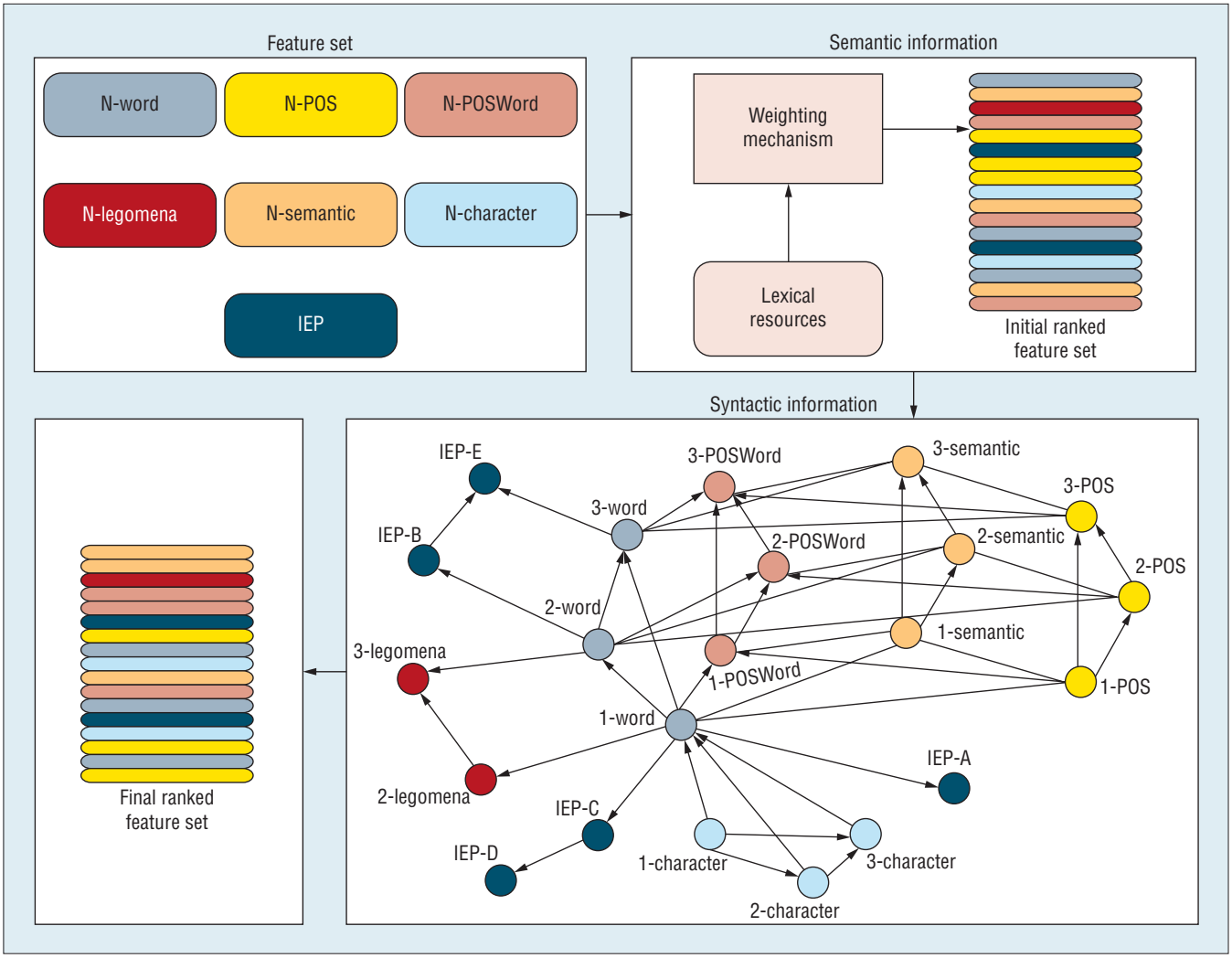

Figure 1. An intelligent feature-selection approach for opinion classification.

An IFS Approach

Figure 1 depicts the design layout for the proposed IFS approach, which uses semantic and syntactic information to refine large input feature spaces. In the example presented here, various categories of n-gram features were used. Although others could also have been incorporated, those utilized include character n-grams, word n-grams, parts-of-speech (POS) tag n-grams, word plus POS tag n-grams, legomena n-grams,[2] information extraction patterns (IEP),[1,3] and semantic patterns. For each category, I use unigrams, bigrams, and trigrams.

Semantic Information

The features' initial weights are an amalgamation of their occurrence distribution across classes in the training data as well as their degree of subjectivity, which is derived from SentiWordNet, a publicly available lexical resource.[6] Figure 2 presents the initial weighting formulation for word n-grams. Given a word n-gram feature a_x that consists of d tokens, the initial weight $w(a_x)$ is the sum of $wt(a_x)$ and $ws(a_x)$, where $ws(a_x)$ is computed by determining the average polarity value across the individual tokens encompassed within the n-gram.

For each token a_{xi}, the polarity value is the average of the sum of its positive and negative scores for each word-sense pair $s(a_{xi}, j)$ in SentiWordNet, where j is one of the k senses of a_{xi}. The computation of $ws(a_x)$ for other n-gram feature categories differs slightly. For instance in the case of parts-of-speech (POS) tag plus word n-grams, the word polarity values are only computed for word-sense pairs in SentiWordNet where the sense has the same POS as that of the tag associated with the word.

Syntactic Information

The IFS approach uses a feature relation network (FRN) that utilizes two important syntactic n-gram relations: subsumption and parallel relations.

In the syntactic information box in Figure 1, subsumption relations are denoted with arrows, while parallel relations are depicted using solid lines. These two relations enable intelligent comparison between features to facilitate enhanced removal of redundant and/or irrelevant attributes. Each remaining feature with a weight greater than 0 is first checked for potential subsumptions, then analyzed for parallel relations.

A subsumption relation occurs between two n-gram feature categories where one category is a more general, lower-order form of the other.[1] A subsumes B (A → B) if B is a higher order n-gram category with n-grams that contain the lower-order n-grams found in A. For example, word unigrams subsume word bigrams and trigrams, while word bigrams subsume word trigrams. Hence, given A → B, we keep features from category B if their weight exceeds that of their general lower-order counterparts found in A by some threshold t.[1] For instance, the bigrams "I love" and "love chocolate" would only be retained if their weight exceeded that of the unigram "love" by t—that is, if they provided additional information over the more general unigram. Otherwise, they would be assigned a final weight of 0.

A parallel relation occurs when two heterogeneous same-order n-gram feature groups may have some features with similar occurrences. For example, word unigrams can be associated with many POS tags, and vice versa. However, certain word and POS tags' occurrences might be highly correlated. Given two n-gram feature groups with potentially correlated attributes, A is considered to be parallel to B (A—B). If two features from categories A and B, respectively, have a correlation coefficient greater than some threshold p, one of the attributes is removed to avoid redundancy—that is, it is assigned a final weight of 0.

The weight for $a_x = (a_{x1}, \ldots a_{xd})$ is

$w(a_x) = wt(a_x) + ws(a_x)$

where $wt(a_x)$ is the weight for feature a_x in the training data, given that v and w are part of the set of c class labels, $v \neq w$, and $c \geq 2$:

$$wt(a_x) = \max_{v,w}\left(P(a_x \mid v)\log\left(\frac{P(a_x \mid v)}{P(a_x \mid w)}\right)\right)$$

and $ws(a_x)$ is the semantic weight for feature a_x:

$$ws(a_x) = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{1}{k}\sum_{j=1}^{k} s(a_{xi}, j)\right)$$

where $s(a_{xi}, j)$ is the sum of the positive and negative scores for the word a_{xi} and j is one of the k senses of a_{xi} in SentiWordNet.

Figure 2. Initial weighting mechanism for word n-grams.

Evaluation

The IFS approach was evaluated on three online product review testbeds, each consisting of 2,000 reviews: digital camera reviews from Epinions, automobile reviews from Edmunds, and movie reviews from Rotten Tomatoes. All three test beds had two classes that were balanced in terms of the number of reviews per class (1,000 each). Five-fold cross validation was used,[7,8] where feature selection was performed on the binary feature presence vectors for the 1,600 training instances during each fold. The selected features were input into a linear kernel support vector machine (SVM) classifier. The 10,000 to 100,000 features with the highest final weights were run in 2,500 feature increments. Hence, 37 feature quantities were used for all three feature sets.

IFS, as well as IFS ablations using only syntactic or semantic information, were compared against two commonly used feature selection methods: information gain and log likelihood. All five of these feature-selection methods were applied to the feature set depicted in Figure 1. Additionally, a word n-gram feature set used in conjunction with log likelihood was also included. Table 1 and Figure 3 shows the evaluation results. Table 1 depicts the area under the

Table 1. Best accuracy and area under the curve (AUC) values for different feature-selection methods across test beds.

	Digital cameras		Automobiles		Movies	
Feature selection	Best accuracy (%)	AUC	Best accuracy (%)	AUC	Best accuracy (%)	AUC
IFS	89.2	1581	90.7	1618	89.7	1582
Semantic IFS	87.8	1566	89.7	1603	88.5	1566
Syntactic IFS	87.6	1559	89.2	1595	87.6	1560
Information gain	86.7	1549	87.8	1574	85.7	1540
Log likelihood	86.1	1540	88.2	1582	85.8	1527
Word n-gram	85.2	1519	86.0	1546	86.0	1539

curve (AUC) value as well as the best percentage accuracy across the different sized feature sets, and Figures 3a through 3c show the accuracies using the top 10,000 to 100,000 features on each of the test beds.

Using semantic and syntactic information, IFS resulted in feature sets with the best accuracy and AUC values on all three test beds. IFS outperformed information gain and log likelihood by 2 to 4 percent in terms of best accuracy and 30 to 55 points in terms of AUC, while the word n-gram feature set was surpassed by 4 to 5 percent in terms of best accuracy. These comparison feature-selection methods were outperformed by the word n-gram feature set on the movie review test bed, demonstrating how larger feature sets can be detrimental when appropriate feature-selection methods are not availed.[1] Moreover, both the semantic and syntactic information contributed to the IFS approach's overall effectiveness, as evidenced by the performance degradation that resulted when either form of information was omitted.

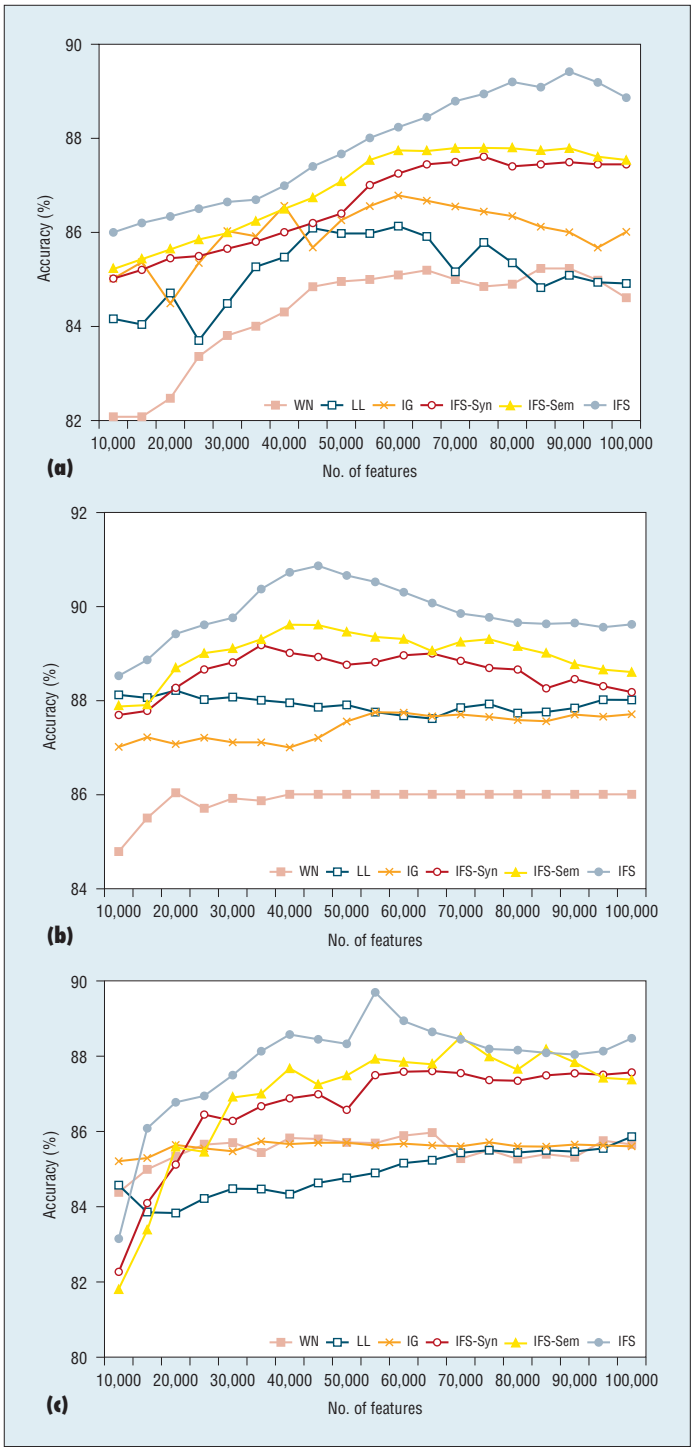

Figure 3. Evaluation results for intelligent feature selection compared to prior feature-selection methods. Online product reviews were tested for (a) digital cameras from Epinions, (b) automobiles from Edmunds, and (c) movie reviews from Rotten Tomatoes.

Future Research

This approach was intended to illustrate how IFS can be combined with larger feature sets for enhanced opinion-classification performance. There are many ways in which IFS for opinion classification can be extended in future research. Numerous additional feature categories could be used, resulting in even more robust feature sets. The syntactic and semantic information modules could be expounded on, for instance, by incorporating additional lexical resources and real-world knowledge bases.

Traditionally, sentiment-analysis research has relied on two types of feature occurrence measures (frequency and presence), while researchers have yet to methodically explore additional distributional and positional measurements. Recently, distributional measures such as compactness and first appearance have been successfully applied to topic-based text categorization.[9] These measures could be used to supplement existing occurrence measures. Hence, we could use IFS mechanisms to reduce opinion-classification feature spaces in a 2D manner: across feature categories (such as specific text features) and various occurrence measures associated with those features.

Future feature-selection efforts could explore the unique challenges associated with performing opinion classification at the document-level versus sentence-, phrase-, and word-level classification. Furthermore, there are other sentiment-analysis tasks that could benefit from improved feature selection, such as opinion holder identification and sentiment target

detection. Given the plethora of potential future directions, one thing is for certain: IFS could help alleviate the quagmire associated with learning features for opinion classification, thereby allowing the kitchen sink to remain where it belongs.

References

1. E. Riloff, J. Wiebe, and T. Wilson, "Learning Subjective Nouns using Extraction Pattern Bootstrapping," *Proc. 7th Conf. Natural Language Learning*, ACM Press, 2003, pp. 25–32.
2. A. Abbasi et al., "Affect Analysis of Web Forums and Blogs using Correlation Ensembles," *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 9, 2008, pp. 1168–1180.
3. E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing*, ACM Press, 2006, pp. 440–448.
4. A. Abbasi and H. Chen, "CyberGate: A Design Framework and System for Text Analysis of Computer Mediated Communication," *MIS Quarterly*, vol. 32, no. 4, 2008, pp. 811–837.
5. Y. Dang, Y. Zhang, and H. Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 46–53.
6. A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," *Proc. 5th Conf. Language Resources and Evaluation* (LREC), European Assoc. Language Resources, 2006, pp. 417–422.
7. T. Mullen and N. Collier, "Sentiment Analysis Using Support Vector Machines with Diverse Information Sources," *Proc. Conf. Empirical Methods in Natural Language Processing*, ACM Press, 2004, pp. 412–418.
8. A. Abbasi et al., "Selecting Attributes for Sentiment Classification using Feature Relation Networks," to be published in *IEEE Trans. Knowledge and Data Eng.*, 2010; http://www.sba.uwm.edu/abbasi_a/index_files/IEEETKDE_FRN.pdf.
9. X.B. Xue and Z.H. Zhou, "Distributional Features for Text Categorization," *IEEE Trans. Knowledge and Data Eng.*, vol. 21, no. 3, 2009, pp. 428–444.

Ahmed Abbasi is an assistant professor of management information systems in the Sheldon B. Lubar School of Business at the University of Wisconsin-Milwaukee. He has a PhD in management information systems from the University of Arizona. Contact him at abbasi@uwm.edu.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.