Markov Blankets and Meta-heuristics Search: Sentiment Extraction from Unstructured Texts

Edoardo Airoldi[1], Xue Bai[1,2,⋆], and Rema Padman[2]

[1] School of Computer Science
Carnegie Mellon University, Pittsburgh, PA USA, 15213
eairoldi@cs.cmu.edu
[2] The John Heinz III School of Public Policy and Management
Carnegie Mellon University, Pittsburgh, PA USA 15213
{xbai, rpadman}@andrew.cmu.edu

Abstract. Extracting sentiments from unstructured text has emerged as an important problem in many disciplines. An accurate method would enable us, for example, to mine online opinions from the Internet and learn customers' preferences for economic or marketing research, or for leveraging a strategic advantage. In this paper, we propose a two-stage Bayesian algorithm that is able to capture the dependencies among words, and, at the same time, finds a vocabulary that is efficient for the purpose of extracting sentiments. Experimental results on online movie reviews and online news show that our algorithm is able to select a parsimonious feature set with substantially fewer predictor variables than in the full data set and leads to better predictions about sentiment orientations than several state-of-the-art machine learning methods. Our findings suggest that sentiments are captured by conditional dependence relations among words, rather than by keywords or high-frequency words.

1 Introduction

Traditionally, researchers have used surveys to collect limited amounts of data in a structured form for their analyses. In recent years, the advent of the Internet, and the widespread use of advanced information technologies in general, have resulted in a surge of information that is freely available online in an *unstructured format*. For example, many discussion groups and review sites exist where people post their opinions about a product. The automatic understanding of *sentiments* expressed within the texts of such posts could lead to a number of new applications in the fields of marketing and information retrieval, and could enable the automated learning of elements of ontologies from online data, e.g., the "trustable level" of the trust ontology [1].

Researchers have been investigating the problem of automatic text categorization for the past two decades. Satisfactory solutions have been found for the cases of topic categorization and of authorship attribution; briefly, topics are captured

[⋆] Corresponding author.

B. Mobasher et al. (Eds.): WebKDD 2004, LNAI 3932, pp. 167–187, 2006.

by sets of keywords, whereas authors are identified by their choices about the use of non-contextual, high-frequency words [2]. Pang et al [3] showed that such solutions, or extensions of them, underperform when ported to sentiment extraction, yielding cross-validated accuracies and areas under the curve (AUC) in the high 70%s to low 80%s. Even more worrisome is the fact that these performances are obtained using large vocabularies, whose words' discriminatory power is likely to be due to chance for many of the words. We conjecture that one reason for the failure of such approaches maybe attributed to the fact that the words used in the classification are *selected independently* of one another, whereas we argue that their very interactions lead to the emergence of sentiments in the text. The goal of this paper is to present a machine learning technique for learning predominant sentiments of online texts, available in unstructured format, that:

– is capable of selecting words that are related to one another and to the sentiment embedded in the texts significantly, i.e., beyond pure chance, and
– is capable of finding a minimal vocabulary that leads to good performance in categorization and prediction tasks.

Our two-stage Markov Blanket Classifier (MBC) learns conditional dependencies among the words and encodes them into a *Markov Blanket Directed Acyclic Graph* (MB DAG) for the sentiment variable (first stage), and then uses a *Tabu Search* (TS) meta-heuristic strategy to fine tune the MB DAG (second stage) in order to yield a higher cross-validated accuracy. Learning dependencies allows us to capture semantic relations and dependent patterns among the words, which help us approximate the meaning of sentences with respect to the sentiment they encode. Further, performing the classification task using a Markov Blanket (MB) for the sentiment variable (in a Bayesian network) has important properties: (a) it specifies a statistically efficient prediction of the probability distribution of the sentiment variable from the smallest subset of predictors, and (b) it provides accuracy while avoiding over-fitting due to redundant predictors. We test our algorithm on the publicly available "movie reviews" data set [4] and on three proprietary corpora of online news with different degrees of topicality [5], and achieve a cross-validated accuracy and AUC comparable to the best performances of competing state-of-the-art classifiers, with an extremely parsimonious vocabulary.

This paper is organized as follows: Section 2 surveys related work. Section 3 provides some background about Bayesian networks, Markov Blankets, and Tabu Search. Section 4 contains details about our proposed methodology. Section 5 describes the data and presents the experimental results. Last, Section 6 discusses of our findings and Section 7 concludes.

2 Related Work on Sentiments

The problem of sentiment extraction is also referred to as opinion extraction or semantic classification in the literature. A related problem is that of studying the semantic orientation, or polarity, of words as defined by Osgood et al. [6].

Hatzivassiloglou and McKeown [7] built a log-linear model to predict the semantic orientation of conjoined adjectives using the conjunctions between them. Huettner and Subasic [8] hand-crafted a cognitive linguistic model for *affection* sentiments based on fuzzy logic. Das and Chen [9] used domain knowledge to manually construct lexicon and grammar rules that aim at capturing the "pulse" of financial markets as expressed by online news about traded stocks. They categorized news as *buy*, *sell* or *neutral* using five classifiers and various voting schemes to achieve an accuracy of 62% (random guesses would top 33%). Turney and Littman [10] proposed a compelling semi-supervised method to learn the polarity of adjectives starting from a small set of adjectives of known polarity, and Turney [11] used this method to predict the opinions of consumers about various objects (movies, cars, banks) and achieved accuracies between 66% and 84%. Pang et al. [3] used off-the-shelf classification methods on frequent, non-contextual words in combination with various heuristics and annotators, and achieved a maximum cross-validated accuracy of 82.9% on data from IMDB. Dave et al. [12] categorized positive versus negative movie reviews using support vector machines on various types of semantic features based on substitutions and proximity, and achieved an accuracy of at most 88.9% on data from Amazon and Cnn.Net. Liu et al. [13] proposed a framework to categorize emotions based on a large dictionary of common sense knowledge and on linguistic models.

3 Problem - Background

We introduce the problem and briefly discuss its scope in Section 3.1. We review concepts relevant to our methodology in Sections 3.2 and 3.3; specifically we review Bayesian networks, Markov blankets, and Tabu search.

3.1 Problem Definition

Our problem can be formally stated as a typical classification problem. Briefly, our data consists of a collection of documents, $\{\ _{d},\ _{d1},\ \ _{dV}\}_{d=1}^{N}$, that is, of examples of the corresponding random variables, $\{\ ,\ _{1},\ \ _{V}\}$. The overall sentiment of each document is encoded by the variable $=\ _{d}$, which can take one of a finite number of sentiment values ($_{d} =\ ,\ = 0,\ ,\ $). The index identifies a unique word in the vocabulary, and denotes the size of the vocabulary observed in the collection of documents. Each of the variables $\{\ _{w}\}_{w=1}^{V}$ encodes the presence or absence of word in document , i.e., $_{dw} \in \{0,1\}$ for $= 1,\ ,\ $.

Problem (Sentiment Classification). *Given a collection of documents, $\{\ _{d1},\ \ _{dV}\}_{d=1}^{N} \in \{0,1\}^{V}$, along with an indication of the corresponding sentiments they encode, $\{\ _{d}\}_{d=1}^{N} \in [0,\ ]$, we want to learn a classifier that predicts the overall sentiments of new documents, $: \{0,1\}^{V} \rightarrow [0,\ ]$, with high accuracy. Documents may show different degrees of topicality.*

Note that, although the mathematical formulation of the problem is that of a typical (supervised) classification problem, its most interesting characteristic is

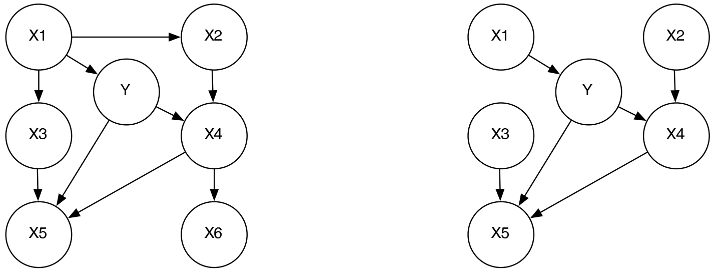

Fig. 1. (left) A sample Bayesian Network (S, P), and (right) the Markov Blanket for the variable encoding the overall sentiment of a document, Y

not expressed by the s and s alone: that is, sentiments are complex semantic elements hardly expressible by mere *independent* presence/absence of words. We develop a methodology to capture *dependent* word presence/absence patterns.

3.2 Bayesian Networks and Markov Blankets

A *Bayesian Network* is a graphical representation of the joint probability distribution of a set of random variables. A Bayesian Network for a set of variables $= \{ {}_1, \; , {}_V\}$ consists of: (i) a directed acyclic graph (DAG) that encodes a set of conditional independence assertions among variables in ; (ii) a set $= \{ {}_1, \; , {}_V\}$ of local conditional probability distributions associated with each node and its parents. A Bayesian Network also has a causal interpretation: a directed edge from one variable to another, $\rightarrow$, represents the claim that is a direct cause of with respect to other variables in the DAG [14,15].

Definition 1. *satisfies the Markov Condition for if every node $_i$ in is independent of its non-descendants and non-parents in , conditional on its parents.*

The Markov Condition implies that the joint distribution can be factorized as a product of conditional probabilities, by specifying the distribution of each node conditional on its parents [15]. In particular, for a given structure , the joint probability distribution for can be written as

$$(\;) = \prod_{i=1}^{V} {}_i({}_i | \; {}_i), \qquad (1)$$

where $_i$ denotes the set of parents of $_i$; this is called a Markov factorization of according to .

Definition 2. *Given the set of variables and target variable , a Markov Blanket (MB) for is the smallest subset of variables in such that is independent of $\setminus$, conditional on the variables in .*

Assuming that there are no conditional independence relations in other than those entailed by the Markov condition for , then for a given Bayesian Network (,), there is a unique Markov Blanket for consisting of ${}_Y$, the set of parents of ; ${}_Y$, the set of children of ; and ${}_Y$, the set of parents of children of .

For example, consider the two DAGs in Figure 1. The factorization of entailed by the Bayesian Network (,) is the following,

$$\begin{aligned}(\,,\, {}_1,\, ,\, {}_6) = {}_Y(\,|\, {}_1)\, {}_4({}_4|\, {}_2,)\, {}_5({}_5|\, {}_3, {}_4,)\, {}_2({}_2|\, {}_1)\times \\ \times\, {}_3({}_3|\, {}_1)\, {}_6({}_6|\, {}_4)\, {}_1({}_1)\end{aligned} \quad (2)$$

Instead, the factorization of the conditional probability entailed by the Markov Blanket for , (| ${}_1$, , ${}_6$), corresponds to the product of those (local) factors in equation 2 that contain the term ,

$$(\,|\, {}_1,\, ,\, {}_6) = {}' \,{}_Y(\,|\, {}_1)\, {}_4({}_4|\, {}_2,)\, {}_5({}_5|\, {}_3, {}_4,), \quad (3)$$

where ${}'$ is a normalizing constant independent of .

Definition 3. *${}_{DAG}$s that entail the same set of conditional independence relations are said to be Markov equivalent; and the set of all ${}_{DAG}$s that are Markov equivalent form a Markov equivalence class.*

3.3 Tabu Search Heuristic

Tabu Search is a meta-heuristic search strategy that is able to guide traditional local search methods to escape the trap of local optimality with the assistance of *adaptive memory* [16]. Tabu search is viewed as "intelligent" search because it makes use of adaptive memory. The adaptive memory feature of TS allows the implementation of procedures that are capable of searching the solution space economically and effectively. In its simplest form, Tabu Search starts with a feasible solution and chooses the *best move* according to an evaluation function while taking steps to ensure that the method does not revisit a solution previously generated. This is accomplished by introducing *tabu restrictions* on possible moves to discourage the reversal and in some cases repetition of selected moves. The *tabu list* that contains these forbidden move attributes is known as the short term memory function. It operates by modifying the search trajectory to exclude moves leading to new solutions that contain attributes (or attribute mixes) belonging to solutions previously visited within a time horizon governed by the short term memory. Intermediate and long-term memory functions may also be incorporated to intensify and diversify the search.

4 Methods: A Markov Blanket for Word Patterns

In this section we describe our methodology and the intuitions behind it. A sketch of the various algorithms described here can be found in Appendix A.

```
**LrnTSMBC** $(\{x_{d1}, ..., x_{dV}\}, \{y_d\}, \delta, \alpha)$

1. $L_Y =$ **Adj** $(\{y_d\}, \{x_{d1}, ..., x_{dV}\}, \delta, \alpha)$
2. **for** $X_i \in L_Y$
   2.1. $L_{X_i} =$ **Adj** $(\{x_{di}\}, \{x_{d1}, ..., x_{dV}\} \backslash x_{di}, \delta, \alpha)$
3. $G =$ **Ornt** $(Y \cup L_Y \cup_i L_{X_i})$
4. $\{MB_{DAG}(Y), L\} =$ **Trsfm** $(G)$
5. **TabuSrch** $(MB_{DAG}(Y), L, Max_{Iter})$
```

Fig. 2. Overview of the algorithm underlying the Tabu-Search-enhanced Markov Blanket Classifier. The relevant parameters are: a data set with V words and N documents; Y, the sentiment variable; δ, the maximum size of separating sets of words considered for the conditional independence tests; α, the significance level for the G^2 statistical independence tests. See the text for more details.

4.1 Markov Blanket Classifier at a Glance

The goal of the *Markov blanket classifier* (MBC) is two-fold: (1) to find a parsimonious vocabulary that is expressive enough to capture the overall sentiment of a document, and (2) to find a dependency structure among words in the vocabulary and the sentiment variable that leads to good predictions of the overall sentiment of a new document. Figure 2 presents an overview of the algorithm that we employ to learn the Tabu-Search-enhanced Markov Blanket Classifier.

The learning algorithm can be divided into two stages. In the **first stage** (steps 1. to 4.), the collection of training documents, , is used to generate an *initial* Markov blanket for the sentiment variable, . The first stage aims at finding a parsimonious, yet expressive vocabulary, and the search for the "right" vocabulary takes into account dependency patterns amongst words. The first stage ends at step 4. with a subset of the words that is meant to be expressive enough to describe word patterns that lead to the emergence of the overall sentiment in the text (the ${}_{DAG}$ for), as well as a list of words that are not part of the Markov Blanket ( ), but which the algorithm cannot exclude (in step 3.) from being useful in expressing sentiments.

However, the initial ${}_{DAG}$ may be highly suboptimal due to the application of repeated conditional independence tests [14], in steps 1. and 2., and propagation of errors in causal orientation [17]. Hence, Tabu Search is applied in the **second stage** (step 5.) to improve the predictive power of the structure of the initial Markov blanket. The algorithm stops after a fixed number of iterations or a fixed number of non-improving iterations.

These two steps are detailed in the rest of this section.

4.2 Description of the Algorithms

In the descriptions of the algorithms that follow, the relevant parameters are: a data set with words and documents; , the sentiment variable; , the maximum size of the separating sets (sets of words) considered for the conditional independence tests; , the significance level for the ^{2} statistical independence tests.[1]

Note that both the ^{2} test of independence and the use of binary variables to encode presence/absence of words are choices dictated by our endeavor for a parsimonious, topic-independent vocabulary. Intuitively, we can divide the words into three groups: common words that are highly frequent in most documents of a collection, topical words that are highly frequent in a few documents, and rare words. Using presence/absence of words, rather than their frequency of occurrence, dampens the discriminative power of topical words, while selecting words with a statistically significant ^{2} statistic favors words that are highly frequent overall. The composite effect is that words that are strongly associated with sentiment are retained. In fact, we can think of sentiments as behaving like "widespread topics" in terms of the frequency pattern of the words that lead to their emergence. Our methodology aims at removing "topical" words associated with narrow topics, while promoting those associated with sentiment, that is, widespread topics. This intuition can be formalized by placing each word on the "frequency spectrum" proposed in [18].

Search for an Initial Markov Blanket DAG. The first stage of *LrnTSMBC* consists of the steps from 1. to 4. in Figure 2. It generates an initial $_{DAG}$ for from the data that reflects dependency patterns among words in the collection of training documents [17].

Searching for Adjacent Nodes. The core of the search for a parsimonious vocabulary that is expressive enough to capture the overall sentiments of documents consists of the steps 1. and 2. There, independence and conditional independence ^{2} tests are carried out according to a breadth first heuristic. In step 1. the function *Adj* is used to identify a list of words (corresponding to nodes of a Bayesian network) related to the sentiment variable, $_Y$, and then in step 2. the function *Adj* is used to identify lists of words related to each word $_i$ in the list $_Y$.

Following the intuition of a Bayesian network, the algorithm starts with a singleton graph containing the target node only. Then it selects those variables among $_1$, , $_V$ that are associated with within a path of length two in the graphical representation, that is, it finds potential parents and children ($_Y$) of , and potential parents and children ($\cup_i$ $_{X_i}$) of nodes $_i \in$ $_Y$, using conditional independence tests.

Through steps 1. and 2., the adjacencies are represented by undirected edges. At the end of step 2. we are left with an undirected graph over the words $\cup$ $_Y \cup_i$ $_{X_i}$, which contains the $_{DAG}$ for in terms of $_Y$, $_Y$, and $_Y$—see Section 3.2 for the precise definition.

Orienting Edges. The algorithm *Ornt* in step 3. is responsible for the orientation of the edges. The edges are oriented by repeatedly applying a set of four edge

[1] Intuitively, G^2 is a variation of the χ^2 test for independence of random variables. For a formal definition see [14].

```
TabuSrch ($MB_{DAG}$, $L$, $Max_{Iter}$)

1. init: $best_{MB} = curr_{MB} = MB_{DAG}$, $best_{Score} = 0$
2. repeat until ($best_{Score}$ does not improve for $k$ consecutive iterations)
   2.1. form $candidate_{Moves}$ for $curr_{MB}$
   2.2. find $best_{Move}$ among $candidate_{Moves}$ according to function score
   2.3. if ($best_{Score} <$ score ($best_{Move}$))
        2.3.1. update $best_{MB}$, by applying $best_{Move}$, and $best_{Score}$
        2.3.2. add $best_{Move}$ to $TabuList$ // not re-considered in the next $m$
               iterations
   2.4. update $current_{MB}$ by applying $best_{Move}$
3. return $best_{MB}$ // an $MB_{DAG}$
```

Fig. 3. Tabu search enhancement

orientation rules [14,19] described in detail in appendix A.2. These rules are meant to recover the true, unobservable directed acyclic graph (DAG) over the selected words. This algorithm provably guarantees to find the correct DAG in the limit (of infinite documents) [17]. The assumption underlying the asymptotic correctness argument is that every conditional independence statement that we can derive from the data also holds in the true graph, which is a typical assumption underlying methods for statistical learning of causal relations [15].

The edge orientation algorithm returns a (possibly) partially oriented DAG. In fact, while the *Ornt* algorithm guarantees the correctness of the result in the limit, it does not guarantee that all edges will be assigned a unique direction in a finite sample.

Forcing a Markov Blanket DAG. The algorithm *Trsfm* at step 4. transforms the output of step 3. into a proper $_{DAG}$ by simply removing the undirected and bi-directed edges, along with the corresponding nodes, that is, words. These words are not thrown away, but are stored in the list ; these are words which were not removed from the battery of conditional independence tests, but for which there is uncertainty as to what role they may have in the $_{DAG}$.

Tabu Search Heuristic. In the final step the algorithm *TabuSrch* is applied to improve the initial $_{DAG}$ and in order to boost the predictive structure of the DAG amongst the words in the selected vocabulary. Our algorithm searches for solutions in the space of logical Markov Blankets, e.g., moves that result in cyclic graphs are not valid. In particular, four kinds of moves are allowed in *TabuSrch*: edge addition, edge deletion, edge reversal, and edge reversal with node pruning, as illustrated in Figure 4.

The algorithm runs for a fixed number of iterations, or until there is no improvement in the scoring criterion for a predetermined number of iterations. At each step, and for each allowed move, the corresponding $_{DAG}$ is computed, its conditional probability factored, its predictions scored, and the best move

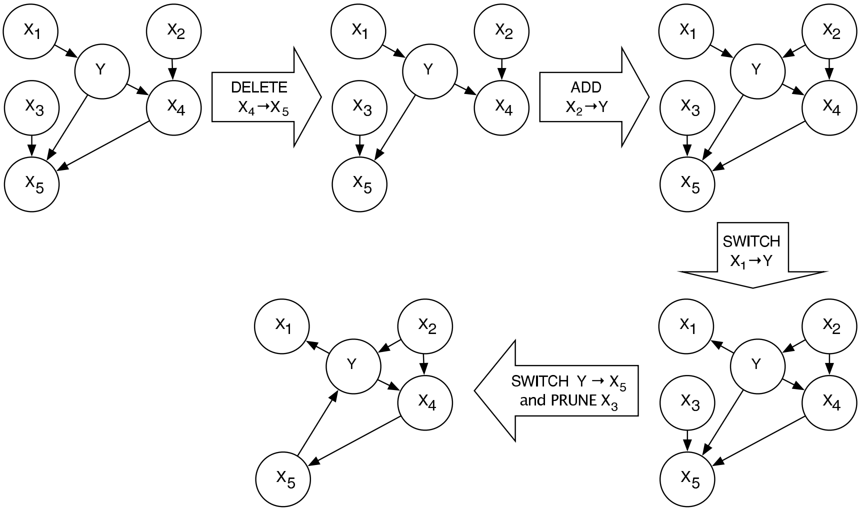

Fig. 4. An example of the moves allowed in Tabu Search

is then selected and applied. The best solution and best score at each step are tracked. The *Tabu list* keeps a record of previous moves, so that moves in the Tabu list will not be repeated until their corresponding *Tabu tenure* expires. The value of m, called the Tabu tenure, is varied according to the complexity of the $_{DAG}$s in different problems. When the dependency structure of the Markov Blanket is very dense, the number of neighboring states that need to be considered may grow exponentially, so a larger Tabu tenure is preferred.

Implementations of simple versions of *TabuSrch* based on Tabu tenures between 7 and 12 have been found to work well in several settings where Tabu restrictions rule out a non-trivial portion of the otherwise available moves [16]. Our setting appears to be of this nature since the structure of the expected $_{DAG}$s is not complex; in our experiments we use a static Tabu tenure of 7.

4.3 Performing the Classification

Once the *LrnTSMBC* algorithm learns the classifier on the training documents, we need to perform the classification of all the documents in the testing set. We approach the classification as a multiple-class classification problem directly, i.e., we do not divide the problem into a series of binary classification problems. The overall sentiment of a new document is assigned according to its posterior probability given the words present in it. Formally, for each new document, $\{\ _1, \quad , \ _V\}$, we compute: $\ _i = \log \left[\frac{P(Y=y_i|\{x_1,\ldots,x_V\})}{P(Y=y_0|\{x_1,\ldots,x_V\})}\right], \forall \ = 1, \quad , \ $, where the $_i$s represent the possible values of the sentiment variable, $\{\ _i\}_{i=0}^{I}$. We choose the sentiment that maximizes the log-odds, $\ ^* = \arg\max \ _i$.

The classification is carried out using logistic regression, whose performance is comparable to more sophisticated methods and has the advantage of being completely automated, i.e., no parameter tuning is necessary [20].

4.4 Theoretical Properties

The proposed Tabu Search Markov Blanket Classifier has two fundamental properties that add “theoretical guarantees” to its good empirical performance.

1. TS-MBC learns the correct Markov blanket in the limit.
2. The complexity of training an TS-MBC is ( ), i.e., linear in the number of documents .

For an in depth analysis of these properties we refer to [17].

5 Experiments

We tested our method on the data set used in Pang et al [3], and on three proprietary collections of online news [5].

5.1 Movie Reviews Data

This data set contains approximately 29,000 posts to the rec.arts.movies.reviews newsgroup archived at the Internet Movie Database (IMDb). The original posts are available in the form of HTML pages. Some pre-processing was performed to produce the version of the data we used. Specifically, only reviews where authors' ratings were expressed explicitly (either by stars or by numerical values) were selected. Then explicit ratings were removed and converted into one of three categories: positive, negative, or neutral. Finally, 700 positive reviews and 700 negative reviews, which the authors of the corpus judged to be more extreme, were selected for our study. Various versions of the data are available online [4].

Pre-Processing. For the purpose of our study, *words* are strings of letters enclosed by non-letters to the left and to the right. Note that our definition excludes punctuation, even though exclamation signs and question marks may be helpful for the task of classifying sentiments. Intuitively the task of sentiment extraction is a hybrid task between authorship attribution and topic categorization; we look for frequent words, possibly not related to the context, that help express lexical patterns, as well as low frequency words which may be specific to few review styles, but very indicative of an opinion. We considered all the words that appeared in more than 8 documents as our input features, whereas words with lower counts were discarded since they are too rare to be helpful in the classification of many reviews. We were left with a total number of 7,716 words, as input features. In our experiments, we represented each document as a vector, $:= [\ _{1},\ ,\ _{7716}]$, of the size of the initial vocabulary, where each $_{i}$ is a binary random variable that takes the value of 1 if the th word in the vocabulary is present in the document and the value of 0 otherwise.

5.2 Online News Data

This collection consists of three sets of 600 news articles each on the following topics: mergers and acquisitions (M&A, 600 documents), finance (600 documents), and mixed news (600 documents). These three corpora have been designed to exhibit increasing levels of specificity; M&A is the most specific corpus, Mixed News is the least specific one, and the news in the Finance corpus falls somewhere in between. Furthermore, the sentiments we consider are three: positive, neutral and negative. Each corpus contains 200 articles of each sentiment category. Articles were manually labeled with a document-level sentiment by three independent annotators from a pool of seven trained annotators; all documents in the corpus have at least a two-way consensus for their sentiment rating. The agreement rate between annotators was found to be consistently above 78%.

Online news aggregators[2] and specialist news sites[3] were used to identify suitable articles for inclusion. Articles were selected to ensure that they were not sarcastic and that they expressed sentiment concerning only one clearly identifiable entity. Articles in the M&A corpus concern only real or speculated mergers, acquisitions, take-overs or joint-ventures. The Finance corpus includes articles concerning all other corporate financial matters, except those that would merit inclusion in the M&A. Finally, the Mixed News corpus contains news and editorial content concerning a broad range of topics. A number of manual inspections of the corpora were conducted to ensure that all the above criteria were met, and to remove any duplicate items.

More in detail, six sentiment categories were initially available for annotation, as follows: (1,2) very positive or very negative: unreservedly or overwhelmingly positive or negative; (3,4) positive or negative: unreservedly, but only mildly positive or negative, or containing mixed sentiment which on balance was positive or negative; (5) neutral: entirely objective, expressing no sentiment; (6) balanced: containing both positive and negative sentiment with no clear bias towards either. All articles that did not have at least a two-way consensus, and those that had a consensus of balanced, were removed. The resulting corpora each contained 200 items of each of the remaining five sentiment categories. To obtain the data used in this paper, the two positive and negative categories were collapsed into a single positive and negative category, by randomly selecting 100 articles in each corpus from each of the four non-neutral sentiment categories.

Pre-Processing. For the purpose of our study, *words* are strings of letters enclosed by non-letters to the left and to the right. Note that our definition excludes punctuation, even though exclamation signs and question marks may be helpful for the task of classifying sentiments. We performed within-topic classification experiments for which we considered all the words in each corpus as the starting pool from which to extract the final vocabulary. This led to starting

[2] E.g. Google News (http://news.google.com).

[3] E.g. Reuters (www.reuters.com), This is Money (www.thisismoney.co.uk),The Motley Fool (www.fool.co.uk) and The Register (www.theregister.com).

Table 1. Characteristics of Online News Data

Data set	Problem	# Words	Examples	Variable Types
Movie Reviews	Positive/Negative	7,716	1,400	Binary
Online News	Positive/Neutral/Negative	8,492	600 × 3	Binary

vocabularies of sizes 11220, 10531, and 15685, for the Finance, M&A and Mixed News corpora respectively.

5.3 Experimental Setup

As mentioned in section 4, the parameters relevant to our experiments were: , the maximum size of the separating sets to consider for conditional independence tests in Adj; and , the significance level of the tests used to decide whether to accept or reject each of these tests. In Table 2 we show the specific values for and which we considered in our experimental design.

In order to estimate the predictive accuracy of our classifier on new documents, we used a five-fold cross-validation scheme for the sentiment classification experiments. In particular, we used a nested cross-validation scheme. That is, we divided each of the five training sets of documents (one for each cross-validation fold, consisting of 4/5 of the documents respectively) into a sub-training and a sub-testing set of documents, at random, using 70% and 30% of the training documents respectively. Steps 1. to 4. would then make use of the sub-training set of documents, whereas step 5. would refine the $_{DAG}$ on the sub-testing set of documents, with the objective of avoiding over-fitting. This process would then be repeated for each pair (,) of the configuration parameters in Table 2.

Table 2. Experimental Parameter Configurations

Performance Measures	Depth of Search δ	Significance Level α
AUC, Accuracy	1, 2, 3	0.001, 0.005, 0.01, 0.05

The dominant configuration of parameters (in terms of accuracy on the sub-testing set of documents) would then be chosen as *the best configuration*. Then, the best configuration $_{DAG}$ for the sentiment variable for a given cross-validation fold, is used to measure the performance on the testing data for that fold.

5.4 Results and Analysis

We compared the performances of our two-stage MB classifier with those of four widely used classifiers: a naïve Bayes classifier based on the multivariate Bernoulli distribution with Laplace prior for unseen words, discussed in Nigam et al. [21], a support vector machine (SVM) classifier, discussed by Joachims [22], an implementation of the voted Perceptron, discussed in Freund and Schapire [23], and a maximum entropy conditional random field learner, introduced by Lafferty et al. [24].

The first two columns of Table 3 compare the two-stage MBC with the performance of the other classifiers using the *whole feature set* as input. More features did not necessarily lead to better performance, as the classifiers were not able to distinguish discriminating words from noise. In such a situation it is reasonable to expect a good performance of the SVM with respect to the other classifiers. As shown in table 3, the two-Stage MB classifier selects 32 relevant words out of 7,716 words in the vocabulary. The feature reduction ratio is 99.58%; the cross-validated AUC based on the 32 words and their dependencies is 87.52%, which is about 5% higher than the best of the other four methods; the corresponding cross-validated accuracy is 78.07%, which is comparable to maximum entropy but less accurate than SVM.[4] The first stage of the MB classifier is able to automatically identify a very discriminating subset of features (or words) that are relevant to the target variable (, the label of the review). Specifically, the selected features are those that form the Markov Blanket for . Other methods need to be paired with a variable selection strategy.

Table 3. Average performances of various classifiers on all words, on the same number of words selected by information gain, and on the same exact words selected by the Markov blanket classifier. These are the results for the movie reviews data. Notes: ** The cross-validated AUC corresponding to the MBC is obtained performing the classification on the subset of 32 selected words, rather than using all the input words.

Input	All Words		32 Words by Information Gain		32 Words by Markov Blanket		
Method	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	# Words Selected
MBC	**87.52****	78.08			**87.52**	**78.08**	**32**
Naïve Bayes	82.61	66.22	**81.46**	72.43	81.81	73.36	
SVM	81.32	**84.07**	67.88	72.21	69.47	73.00	
Voted perc.	77.09	70.00	78.68	71.71	80.61	73.93	
Max. entropy	75.79	79.43	69.11	**72.86**	69.81	73.44	

At this point, however, the source of the differential in the observed accuracies and AUC is not clear. A possibility could be that the competing methods perform better on an optimized subset of words as it is the case for the two stages of MBC classifier—we expect this to be not true for the SVM though, which typically benefits from a large amount of predictors. To investigate this point, we conducted two additional sets of experiments for the competing classifiers. The corresponding results are reported in columns three to six. Both sets of the experiments optimize the vocabulary that is used to perform predictions

[4] Current SVM implementations are very sensitive to the capacity parameter.

Table 4. Average performances of various classifiers on all words, on the same number of words selected by information gain, and on the same exact words selected by the Markov blanket classifier. These are the results for the financial, mergers & acquisitions and mixed online news data, from top to bottom respectively. Notes: ** The cross-validated accuracy corresponding to the MBC is obtained performing the classification on the subset of 36 (43) selected words, rather than using all the input words.

Input	All Words		36(43) Words by Information Gain		36(43) Words by Markov Blanket		
Method	Kappa (%)	Accuracy (%)	Kappa (%)	Accuracy (%)	Kappa (%)	Accuracy (%)	# Selected Words
MBC	N/A	73.21**			N/A	**73.21**	**36**
Naïve Bayes	41.99	61.33	10.00	40.00	23.75	49.16	
Poisson	55.25	70.16	38.00	58.66	44.99	63.33	
Voted perc.	0.75	33.83	19.75	46.50	22.25	48.16	
Max. entropy	**59.99**	**73.33**	**54.50**	**69.66**	**59.00**	72.66	
MBC	N/A	75.32**			N/A	**75.32**	**43**
Naïve Bayes	44.00	62.66	9.50	39.70	27.50	51.66	
Poisson	59.75	73.16	46.75	64.50	48.75	65.84	
Voted perc.	7.00	38.00	7.75	38.50	16.75	44.50	
Max. entropy	**66.25**	**77.50**	**58.25**	**72.16**	**58.25**	72.16	
MBC	N/A	76.68**			N/A	**76.68**	**43**
Naïve Bayes	47.75	65.16	12.25	41.50	28.50	52.33	
Poisson	57.75	71.84	39.00	59.33	44.99	63.33	
Voted perc.	1.50	34.33	9.25	39.50	10.50	40.33	
Max. entropy	**71.25**	**80.84**	**53.49**	**69.00**	**62.50**	75.00	

about sentiments using a feature selection criterion first, then feed optimized vocabulary into the competing classifiers as predictor variables. In particular, the first set of results is obtained by making predictions on words selected by information gain. The second set of results is obtained by making predictions using the words selected by MBC. A comparison between performances of competing classifiers on vocabularies by information gain selection criterion versus those by MBC tells us about the differential effect of MBC as a word selection criterion. A comparison between performances of MBC and those of competing classifiers, both on the full set of words and on two version of optimized vocabularies tells us about the differential effect of MBC as a classifier.

Columns three and four in Table 3 compare the performance of the two-stage MBC with others classifiers using the *same number of features* selected by information gain. In fact, information gain allows us to rank the features from most to least discriminating but gives no indication of how many words are correlated with the sentiment variable, significantly beyond chance. In this second comparison the two-Stage MB classifier dominates the other methods both in terms of AUC and accuracy, though it is not clear (yet) whether the extra performance comes form the different feature selection strategies, or from the dependencies encoded by the MB. To investigate this point, columns five and six in Table 3 compare the performance of the two-stage MBC with others classifiers using the *same exact features*. We find that the set of words selected as part of the Markov blanket for the sentiment variable contains better discriminative words, in fact all the four competing classifiers performed better on the set of features in the Markov blanket. Further, these words are also significantly discriminative beyond pure chance.

We repeated the same battery of experiments on the online news corpora and obtained similar results. Notice that AUC is for binary class problems only thus we report the Kappa statistics instead, which intuitively measures the incremental performance of a classifier with respect to the baseline accuracy[5]. The baseline accuracy is the accuracy of randomly guessing the sentiment class, and in the three class problem the expected accuracy is 33.33%. We did not report the results for the SVM because a multi-class implementation was not readily available to us, we reported instead the accuracy of a Poisson classifier [26,18].

These experiments are in line with more results we have obtained on several medical data sets [27]. Further, according to the empirical findings of Pang et al [3], the baseline accuracy for human-selected vocabularies can be set at about 70%. Comparing the human intuition to our fully automated machine learning technique (two-stage MBC), we observe a non-negligible improvement.

6 Discussion

The main findings that emerge from our experiments are as follows.

1. The problem of learning sentiments is essentially a word selection problem.

Pairwise dependencies amongst words, and dependent patterns, play a crucial role in finding a parsimonious, yet expressive vocabulary. The two-stage MB classifier leads to more robust predictions by: (i) selecting statistically discriminating words with respect to the sentiment variable, and (ii) learning those dependencies among the words that lead to the emergence of sentiments in the texts. Although the Tabu heuristic search in the second stage may return a final DAG where few dependencies are present amongst words, the initial DAG always includes a rich dependency structure.

2. The accuracy of alternative methods can be improved by using the parsimonious vocabulary selected in the first stage of TS-MBC.

Tables 3 and 4 present the performance of several widely used classifiers in extracting sentiments from movie reviews (two sentiments) and online news (three

[5] The kappa statistic was introduced by Carletta [25] to assess the quality of a classifier in terms of accuracy above a random baseline. It is defined as $\kappa = \frac{A-R}{1-R}$, where A is the empirical probability of agreement on a category, and R is the probability of agreement for two annotators that label documents at random (with the empirically observed frequency of each label). Hence kappa ranges from -1 to $+1$; positive values indicate a performance better than the random baseline, whereas negative values indicate a worse performance.

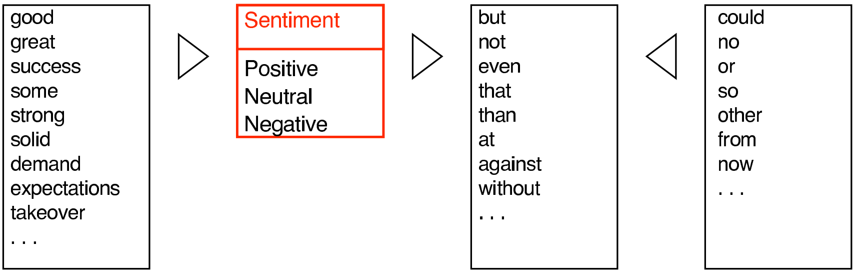

Fig. 5. An example of the words in the final vocabulary and their connections

sentiments). Using the words selected by the TS-MBC as predictors improves both cross-validated accuracy and AUC for all classifiers. See [28] for more details.

3. Words that are indicative of sentiments are not "rare."

The starting pools of words corresponding to online news data contained ten to 15 thousand words. A sample of the words we used to perform the experiments in Tables 3 and 4 is shown in Figure 5. Most of the words appear more than eight times every ten thousand words, and several of them are non-contextual, high-frequency words.

4. TS-MBC tends to select words that are relevant and "exportable."

Words that are intuitively important in predicting sentiments, e.g. *important, success, positive, solid, expected, strong*, as well as pair of non-contextual words, e.g., *no-but, or-not, POS-but, NEG-but* and others (e.g., see Figure 5), were selected in the final vocabulary by the TS-MBC procedure. These same words were not deemed as important for predicting sentiments according to their information gain score, which may be one of the reasons behind the poorer performance entailed by the words selected according to it.

Furthermore, as we discussed in 4.2 in more detail, both the ^{2} test of independence and the use of binary variables to encode presence/absence of words are choices suited to creating a parsimonious, "topic-independent" vocabulary. For example, words that capture linguistic issues such as negation, ifs, buts, howevers, and comparisons appear in the selected vocabularies. Our methodology aims at removing words associated with latent concepts that span few documents, i.e., topics, while promoting those words associated with latent concepts that span a large number of documents, i.e., sentiments.[6].

In our experiments, however, we observed only a modest overlap between the selected vocabularies for the three data sets. Given the small number of texts in our data sets, it is possible that several different subsets of a small number of words may lead to very good classification results (i.e., presence of several local

[6] This intuition can be formalized by placing each word on the "frequency spectrum" proposed in [18].

optima), and that the TS-MBC may be finding a few of them, but different ones for different collections. More experiments on a larger number of corpora, and on more documents, are needed in order to draw stronger conclusions.

5. Tabu search improves the prediction performance.
6. The prediction performance is consistent on two and three class problems.

We note that our investigations generalize previous attempts particularly because we are working with both two-class and three-class problems, whereas previous works have explored two-class problems. Further, our results are consistent on both problem settings.

7 Conclusions

In this paper we have proposed the Tabu-search-enhanced Markov blanket classifier (TS-MBC) and we have shown that: (1) it is a fully automated system able to select a parsimonious vocabulary, customized for the classification task at hand in terms of size and relevant features; (2) it is able to capture and take advantage of dependencies among words while selecting the vocabulary; (3) it learns the correct Markov blanket in the limit and its complexity is linear in the number of training documents.

Our experiments show that the problem of sentiment classification is essentially a word selection problem, and our findings suggest that words that occur often, along with their statistical dependencies and few strong adjectives, include most of the vocabulary needed to express sentiments and perform reasonable predictions. As a feature selection method, our TS-MBC leads to vocabularies that enhance the predictive performance of several popular classifiers with respect to vocabularies selected with information gain. When paired with logistic regression to perform classification, our TS-MBC leads to predictive performance comparable to that of state-of-the-art classification methods. Most importantly, the limited size of the vocabulary allows for interpretability and re-usability.

In conclusion, we believe that in order to capture sentiments we have to move beyond the search for richer feature sets and the independence assumption. Rather it is crucial to capture those dependencies amongst words that lead to the emergence of context and meaning.

7.1 Future Work

In future work we plan to investigate: (1) strategies that allow TS-MBC to reach close local optima starting from corpora with different degree of topicality, i.e., strategies to select overlapping, more consistent vocabularies for predicting sentiments; (2) the performance with larger data sets, more sentiment categories, and on shorter texts, e.g., sentences; (3) cross-genre transferability; (4) the effect of pre-processing the starting pool of words, e.g., through the use of stemming, POS tagging, lemmatization or thesauri, as well as the use of generalized features for non-word, e.g., numbers, dates and punctuation; (5) how performance varies when reducing the size of the initial vocabulary, selected by frequency in corpus.

Acknowledgments

The authors wish to thank William Cohen, Clark Glymour, and Peter Spirtes, at the School of Computer Science of Carnegie Mellon University, for helpful discussions and insights; Joseph Ramsey, for helpful discussions and support during the implementation of the algorithms; Lillian Lee and Bo Pang at Cornell University for helpful discussions and suggestions; Roy Lipski for helpful suggestions and for the sharing the online news data, prepared by Infonic Ltd.; and the ACM SIGKDD anonymous reviewers for helpful comments.

References

1. Golbeck, J., Hendler, J.: Accuracy of metrics for inferring trust and reputation. In: Proceedings of 14th International Conference on Knowledge Engineering and Knowledge Management. (2004)
2. Airoldi, E.M., Anderson, A.G., Fienberg, S.E., Skinner, K.K.: Who wrote Ronald Reagan radio addresses? Bayesian Analysis **1** (2006) 289–320
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. (2002) 79–86
4. Online movie reviews data. http://www.cs.cornell.edu/people/pabo/movie-review-data/.
5. Online news data. http://www.infonic.com/.
6. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois Press, Chicago, Illinois (1957)
7. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, ACL (1997) 174–181
8. Huettner, A., Subasic, P.: Fuzzy typing for document management. In: Association for Computational Linguistics 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes. (2000) 26–27
9. Das, S., Chen, M.: Yahoo! for amazon: Sentiment parsing from small talk on the web. In: Proceedings of the Eighth Asia Pacific Finance Association Annual Conference, APFA (2001)
10. Turney, P., Littman, M.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council, Canada (2002)
11. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings Fortieth Annual Meeting of the Association for Computational Linguistics. (2002) 417–424
12. Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: roceedings of the Twelfth International Conference on World Wide Web. (2003) 519–528

13. Liu, H., Lieberman, H., Selker, T.: A model of textual affect sensing using real-world knowledge. In: Proceedings of the Eighth International Conference on Intelligent User Interfaces. (2003) 125–132
14. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT Press (2000)
15. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
16. Glover, F.: Tabu Search. Kluwer Academic Publishers (1997)
17. Bai, X.: Tabu search enhanced graphical models for classification of high dimensional data. Technical Report CMU-CALD-05-101, School of Computer Science, Carnegie Mellon University (2005)
18. Airoldi, E., Cohen, W., Fienberg, S.: Bayesian models for frequent terms in text. Manuscript (2005)
19. Spirtes, P., Meek, C.: Learning bayesian networks with discrete variables from data. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press (1995) 294–299
20. Komarek, P., Moore, A.: Making logistic regression a core data mining tool (2005) Manuscript.
21. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Machine Learning **39** (2000) 103–134
22. Joachims, T.: A statistical learning model of text classification with support vector machines. In: Proceedings of the Conference on Research and Development in Information Retrieval, ACM (2001) 128–136
23. Freund, Y., Schapire, R.: Large margin classification using the perceptron algorithm. Machine Learning **37** (1999) 277–296
24. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. (2001) 282–289
25. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22** (1996) 249–254
26. Airoldi, E., Anderson, A., Fienberg, S., Skinner, K.: Who wrote Ronald Reagan radio addresses? Journal of Bayesian Analysis (2005) to appear.
27. Ramsey, J., Bai, X., Glymour, C., Padman, R., Spirtis, P.: Mb fan search classifier for large data sets with few cases. Working paper, Department of Philosophy, Carnegie Mellon University (2004)
28. Bai, X., Padman, R., Airoldi, E.: Sentiment extraction from unstructured text using tabu search-enhanced markov blanket. In: Proceedings of KDD Workshop on Mining for and from the Semantic Web (MSWKDD). (2004)
29. Cohen, W.: Minor-third: Methods for identi-fying names and ontological relations in text using heuristics for inducing regularities from data. http://minorthird.sourceforge.net (2004)
30. Bishop, Y., Fienberg, S., Holland, P.: Discrete Multivariate Analysis. Theory and practice. MIT Press (1975)
31. Chickering, D., Meek, C., Heckerman, D.: Large-sample learning of bayesian networks is np-hard. In: Proceedings of Nineteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann (2003) 124–133
32. Engstrom, C.: Topic dependence in sentiment classification. Technical Report 07-22-2004, St Edmunds College, University of Cambridge (2004)
33. Finn, A., Kushmerick, N.: Learning to classify documents according to genre. In: IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis. (2003)

34. Koller, D., Sahami, M.: Towards optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann (1996) 284–292
35. Lewis, D.D.: Evaluating Text Categorization. In: Proceedings of Speech and Natural Language Workshop, Morgan Kaufmann (1991) 312–318
36. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Advances in Neural Information Processing System. (1999)
37. Mitchell, T.: Machine Learning. McGraw-Hill (1997)
38. Montgomery, A., Kannan, S.: Learning about customers without asking. GSIA Working Paper,Carnegie Mellon University (2002)
39. Piatetsky-Shapiro, G., Steingold, S.: Measuring lift quality in database marketing. SIGKDD Explorations, **2** (2000) 7680
40. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning. (1998) 445–453
41. E.P. Xing, M.J., Karp, R.: Feature selection for high-dimensional genomic microarray data. In: Proceedings 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001)

A Sketch of the Algorithms

Below we present a sketch of the algorithms used in 4.2.

A.1 Searching for Adjacent Nodes

```
Adj (Node $Y$, Node List $L$, Depth $\delta$, Significance $\alpha$)

1. $A_Y := \{X_i \in L$: $X_i$ is dependent of $Y$ at level $\alpha\}$
2. for $X_i \in A_Y$ and for all distinct subsets $S \subset \{A_Y \backslash X_i\}^d$
   2.1. if $X_i$ is independent of $Y$ given $S$ at level $\alpha$
   2.2. then remove $X_i$ from $A_Y$
3. for $X_i \in A_Y$
   3.1. $A_{X_i} := \{X_j \in L$: $X_j$ is dependent of $X_i$ at level $\alpha$, $j \neq i\}$
   3.2. for all distinct subsets $S \subset \{A_{X_i}\}^d$
      3.2.1. if $X_i$ is independent of $Y$ given $S$ at level $\alpha$
      3.2.2. then remove $X_i$ from $A_Y$
4. return $A_Y$
```

The conditional independence tests are meant to test whether a pair of words is independent conditionally on a set of words. The set of words in the conditional statement is called the separating set, denoted as (,), and formally defined as a mapping of a set of nodes s.t. ($\perp$ | (,)).

A.2 Orienting Edges

```
Ornt (Graph $G$)

Apply the following 4 rules iteratively wherever it applies:

1. Rule 1 (Collider Orientation Rule): for each triple of vertices $(X, V, Z)$
   in $G$, if pair $(X, V)$ and $(V, Z)$ are adjacent, pair $(X, Z)$ are not adjacent
   (i.e. a pattern: $X - V - Z$ ), and if $V \notin sepSet(X, Z)$, then orient $X - V - Z$
   as $X \rightarrow V \leftarrow Z$.
2. Rule 2: for each triple of vertices $(X, V, Z)$ in $G$, if $X \rightarrow V$ and $V$, $Z$ are
   adjacent, $X$, $Z$ are not adjacent (i.e. a pattern $X \rightarrow V - Z$), and there is no
   arrow into $V$, then orient $V - Z$ as $V \rightarrow Z$.
3. Rule 3: if $(X \rightarrow Z \rightarrow V)$, and $\exists$ (undirected edge between $X$ and $V$, i.e.
   pattern $X - V$), then orient $X - V$ as$X \rightarrow V$
4. Rule 4: for any undirected edge connected to $X$ (i.e.$X - V$), if $\exists$ $(Z, W)$
   s.t. $Z$ adjacent to $X$, $W$ is adjacent to $X$, $Z$ is not adjacent to $W$, and there
   is a pattern $W \rightarrow V \leftarrow Z$, then orient $X - V$ as $X \leftarrow V$
```

A.3 Forcing a Markov Blanket DAG

```
Trsfm (Graph $G$, Target $Y$)

1. for any $X$ in $G$ s. t. $X \leftrightarrow Y$ or $X - Y$ , reorient this edge as $X \rightarrow Y$
2. for any $X, Z$ in $G$ s. t. $Z - X \rightarrow Y$ $(Z \notin \{Y\} \cup ParentChild(Y))$, remove
   the edge $Z - X$
3. for any $X$ , $X \notin \{Y\} \cup ParentChild(Y) \cup ParentChild(ParentChild(Y))$
   1.1. remove $X$ and all the associated edges
4. return $G$
```