문자열 커널을 이용한 인터넷 영화평의 감정 분석

김상도○ 윤희근 박성배 박세영 이상조
경북대학교 컴퓨터공학과
{sdkim○, hkyoon, sbpark, sypark sjlee}@sejong.knu.ac.kr

A Sentiment Analysis of Internet Movie Reviews Using String Kernels

Sang-Do Kim○ Hee-Geun Yoon Seong-Bae Park Se-Young Park Sang-Jo Lee
Department of Computer Engineering
Kyungpook National University

요 약

오늘날 인터넷은 개인의 감정, 의견을 서로 공유할 수 있는 공간이 되고 있다. 하지만 인터넷에는 너무나 방대한 문서가 존재하기 때문에 다른 사용자들의 감정, 의견 정보를 개인의 의사 결정에 활용하기가 쉽지 않다. 최근 들어 감정이나 의견을 자동으로 추출하기 위한 연구가 활발하게 진행되고 있으며, 감정 분석에 관한 기존 연구들은 대부분 어구의 극성(polarity) 정보가 있는 감정 사전을 사용하고 있다. 하지만 인터넷에는 나날이 신조어가 새로 생기고 언어 파괴 현상이 자주 일어나기 때문에 사전에 기반한 방법은 한계가 있다. 본 논문은 감정 분석 문제를 긍정과 부정으로 구분하는 이진 분류 문제로 본다. 이진 분류 문제에서 탁월한 성능을 보이는 Support Vector Machines(SVM)을 사용하며, 문서들 간의 유사도 계산을 위해 문장의 부분 문자열을 비교하는 문자열 커널을 사용한다. 실험 결과, 실제 영화평에서 제안된 모델이 비교 대상으로 삼은 Bag of Words(BOW) 모델보다 안정적인 성능을 보였다.

주제어: 영화 평가, 감정 분석, 문자열 커널, Support Vector Machines(SVM)

1. 서 론

다양한 블로그, 커뮤니티 사이트의 등장으로 컴퓨터에 능통하지 않은 사용자들도 자신의 의견을 손쉽게 공유할 수 있게 되었다. 이를 바탕으로 오늘날의 인터넷은 단순히 지식의 공유뿐만 아니라, 사용자들의 감정, 의견 등 다양한 정보들을 공유하는 공간으로 활용되고 있다. 이런 인터넷의 변화로 인해 사용자들은 상품을 구매하기 위하여 다른 사용자들의 의견을 참조하거나 또는 기업에서 소비자들의 반응을 파악하고, 새로운 마케팅 전략의 수립 등, 다양한 용도로 사용되고 있다.

위와 같은 정보들을 활용하기 위하여 인터넷의 각종 문서에서 의견이나 감정의 파악을 위한 수요가 증대되고 있다. 하지만 오늘날의 인터넷에는 너무나 방대한 자료가 존재하기 때문에 사용자가 필요로 하는 문서를 일일이 확인하여 찾는 것은 불가능하다. 이런 이유로 문서에서 감정 또는 의견을 자동으로 추출하기 위한 연구가 활발하게 진행되고 있다.

감정 분석(sentiment analysis)이란 자연어 처리(NLP) 기술을 활용하여 의견이 표현된 문서, 문장 또는 구(phrase) 단위의 극성을 파악하는 연구이다. 이런 연구를 통하여 감정 분석의 극성에 따라 문서들을 분류하고, 사용자의 요구에 따라 다양한 정보를 효율적으로 제공해 줄 수 있다.

영어권에서는 SentiWordNet[1]과 같은 사전 기반의 감정 분석 기술이 제안되었다. 각 어구의 극성 정보를 부착해 놓은 사전을 이용하여 문장의 극성을 판단한다. 그러나 사전 기반 방법을 한국어에 적용하기 위해서는 두 가지의 문제점이 있다. 한 가지는 아직까지 한국어에는 의견성 어구에 관련된 적당한 리소스가 없다는 점이다. 사전을 구축하기 위해서는 매우 많은 비용이 들기 때문에 이런 방법을 감정 분석에 적용하기가 쉽지 않다. 이를 해결하기 위하여 기존의 영어권 리소스를 번역하여 한국어 감정 사전을 구축하고자 하는 시도도 있었으나 타 방법에 비해 큰 성능 향상을 이끌어낼 순 없었다[2].

또 다른 문제는 어구 단위의 사전으로는 극성의 모호성을 해결할 수 없다는 것이다. 이 문제는 한국어뿐만 아니라 타 언어에서도 나타날 수 있다. '영화관에서 정석을 풀게 해준 고마운 영화' 라는 문장을 예로 들어보자. 일반적으로 '고마운' 은 긍정적인 어감을 갖는 단어이다. 하지만 앞에 '정석을 풀게 해준' 이라는 문맥과 결합됨으로 인해서, 부정적인 느낌을 가지게 된다. 이러한 문제는 미리 구축해 놓은 어구 사전으로는 해결할 수 없는 문제이다.

이런 문제를 해결하기 위하여 감정 사전을 사용하지 않은 연구가 있었다[3]. 김묘실과 강승식은 형태소 분석의 결과를 기반으로 한 BOW 모델을 사용하였다[4]. 하

지만 이 모델은 올바른 문장이 쓰여 있는 문서를 분류하기에는 적합하나, 인터넷에 존재하는 문서를 분석하기에는 적합하지 않다. 이는 일반적으로 인터넷 문서에는 언어파괴 현상이 빈번하기 때문에 형태소 분석기를 통해 올바른 분석을 수행할 수 없기 때문이다.

본 논문에서는 기존 방법들의 문제점을 해결하기 위한 새로운 감정 분석 방법을 제안한다. 본 방법에서는 인터넷 게시판에 작성된 의견 문서를 긍정과 부정의 극성으로 나누는 이진 분류 문제로 본다. 실험에서 사용한 분류기는 이진 분류 문제에서 좋은 성능을 보이는 SVM을 이용한다. 각 문서간의 유사도는 SentiWordNet과 같은 추가적인 외부 리소스를 이용하지 않고, 문자열의 부분 문자열만을 비교하는 문자열 커널[5]을 이용하여 구한다. 제안한 모델의 성능 평가를 위하여 본 방법을 자연어 처리 분야에서 문서 분류를 위해 일반적으로 사용하는 BOW 모델과 비교하였다.

논문의 구성은 다음과 같다. 2장에서는 감정 분석 문제를 풀고자한 기존 연구들에 대해 살펴보고, 3장에서는 제안하는 알고리즘에 대해서 설명한다. 4장에서는 실험에 사용된 데이터 및 결과에 대해서 평가하고, 5장에서 결론을 기술한다.

2. 관련 연구

최근에는 상업적 용도로 사용하기 위한 온라인 감정 분석 연구가 급속하게 증가하고 있다. 사용자들은 상품의 평가를 게시하는데 큰 관심을 가지고 있고 웹에는 이런 사용자들을 위한 피드백 지원 시스템이 많이 있다. Hu는 이러한 시스템을 기반으로 디지털 카메라의 크기와 같이 다양한 상품의 특성에 긍정과 부정 의견인 온라인 문서를 분석 하였다[6].

Turney는 특정 패턴을 갖는 의견성 구(phrase)에 PMI(Pointwise Mutual Information)를 이용하여 구에 극성을 할당하고 이를 바탕으로 문서의 극성을 분류하였다[7]. PMI는 비슷한 성질의 어구는 함께 나타나는 빈도가 높을 것이므로 이러한 가정에 근거하여 단어 사이의 관계를 측정한다. Turney는 극성의 방향이 확연히 드러나는 긍정 어휘와 의견성 구의 PMI에서 부정 어휘와 의견성 구의 PMI 지수의 차이를 기준으로 의견성 구에 극성을 할당하였다. 또 다른 연구에서는 문서를 긍정 또는 부정으로 나누는 이진 분류 문제로 보고 기계학습 기법을 적용하였다[8,9].

Paula는 Wikitionary 등의 어휘 사전 리소스를 사용하여 극성이 분명히 드러나는 형용사의 씨앗 목록(seed list)을 확장하여 문서에서 포함된 모든 형용사에 극성을 판별하여 문서를 분류하였다[10]. 이는 저자가 의견을 나타내는데 있어서, 특정 품사의 사용을 중요하게 본 것으로, 일반적인 상식에서 형용사가 많이 사용된 문서는 그렇지 않은 문서보다 주관적일 것이라는 가정을 기반으로 하고 있다.

국내에는 영어권 사전 리소스를 사용하여 한국어 시소러스 구축하여 의견 어구를 분류하는 모델을 제안하는 연구들이 있었다[2,11,12]. 다른 연구에서는 웹에서 얻을 수 있는 리뷰 문장 데이터만 가지고 의견성 어구 사전 없이 의견성 어구에 극성을 할당하였다[3]. 김묘실과 강승식은 문서를 형태소 분석을 바탕으로 자질 벡터(feature vector)를 추출하고, SVM을 사용하여 문서를 판별하였다[4].

3. 감정 분류 시스템

3.1 SVM

본 연구에서는 감정 문서 분류기로 SVM을 사용하였다. SVM은 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년 Vapnik에 의해 소개된 학습 기법으로, 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 초평면(optimal hyperplane)을 찾는 모델이다. SVM은 두 클래스의 경계면으로 마진(margin)을 가장 최대화하는 초평면을 찾는다. 그림 1은 하나의 초평면에 의해서 정해지는 마진을 보여준다.

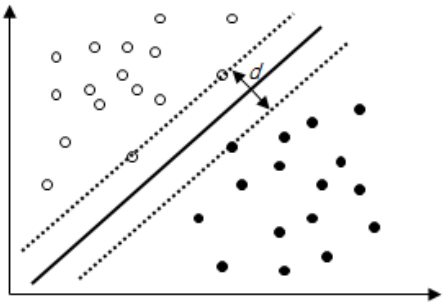

그림 1. 초평면의 마진

SVM에서의 초평면은 식 (1)과 같이 나타낼 수 있다.

$$\vec{w} \cdot \vec{x} + b = 0 \quad (1)$$

여기서 $\vec{x}$는 분류하고자 하는 문서 벡터이며 $\vec{w}$와 b는 학습 데이터로부터 결정되는 파라미터이다. 학습 문서 집합을 $D=\{(y_i, \vec{x_i})\}$과 같이 나타냈을 때, 각각의 학습

문서 벡터($\vec{x_i}$)가 긍정 클래스에 속한 문서이면 y_i의 값에 $+1$을 할당하고, 부정 클래스에 속하는 문서에는 -1을 할당한다. SVM은 식 (2)과 식 (3)을 만족시키는 $\vec{w}$와 b를 찾는 문제이다[13].

$$\vec{w} \cdot \vec{x} - b \geq +1 \quad When \;\; y_i = +1 \tag{2}$$
$$\vec{w} \cdot \vec{x} - b \geq -1 \quad When \;\; y_i = -1 \tag{3}$$

SVM은 선형 문제(linearly separable problem)에 사용되는 알고리즘이지만, 비선형 문제에도 적용할 수 있다. 즉, 비선형 공간의 데이터를 고차원 선형 공간으로 맵핑함으로써 비선형 문제를 선형 문제로 바꾸어 해결할 수 있다. SVM에서는 커널 함수를 사용하여 고차원 선형 공간상에서의 데이터 유사도를 측정한다.

3.2 문자열 커널

문자열 커널은 Lodhi et al.[5]이 제안한 것으로 기존의 문서를 벡터로 나타내기 위해 일반적으로 사용되던 BOW 과 달리, 문서에 포함되어 있는 부분 문자열을 기반으로 하는 커널이다. 문자열 커널은 특정 문자열의 자질을 해당 문자열에 속해 있는 부분 문자열들의 출현 횟수로 나타낸다.

문서에서 나올 수 있는 문자의 집합을 $\sum$라고 할 때, $\sum$에 속한 문자를 유한하게 배열한 문자열 s를 자질 공간으로 사상시키는 함수 Φ는 다음과 같이 정의된다.

$$\Phi_u(s) = \sum_{i:\, u = s[i]} \lambda^{l(i)}$$

여기서 u는 문자열 s에 속한 부분 문자열이다. $i = (i_1, ..., i_{|u|})$는 $1 \leq i_1 < ... < i_{|u|} \leq |s|$을 만족하는, 부분 문자열 u을 구성하고 있는 문자들의 인덱스 집합을 의미한다. λ는 1보다 작은 상수이다. $l(i)$는 인덱스 i의 길이를 의미하며, 이는 $i_{|u|} - i_1 + 1$로 계산된다. 이 자질 벡터를 기반으로 두 개의 문자열 s와 t의 내적은 다음과 같이 계산된다.

$$K_n(s,t) = \sum_{u \in \sum^n} \Phi_u(s)\Phi_u(t) = \sum_{u \in \sum^n} \sum_{i:\, u = s[i]} \sum_{j:\, u = s[j]} \lambda^{l(i)+l(j)}$$

여기서 Φ는 부분 문자열의 길이 n이 4보다 커지면 계산량이 너무 많아 자질 벡터의 값을 직접적으로 계산하기가 힘들다[5]. 하지만 문자열 커널에서는 다음과 같은 새로운 함수를 도입하여 이를 효율적으로 계산할 수가 있다.

$$K'_i(s,t) = \sum_{u \in \sum^n} \sum_{i:\, u = s[i]} \sum_{j:\, u = s[j]} \lambda^{|s|+|t|+t_i+j_1+2}$$

위의 함수는 $l(i)$와 $l(j)$ 대신에 특정 문자열의 시작 위치에서 각 문자열 끝까지의 길이를 사용한다. 이 함수는 다음과 같은 재귀적인 규칙을 통하여 계산되며, 이를 바탕으로 두 문자열 s와 t의 내적 $K_n(s,t)$는 다음과 같이 계산할 수 있다.

$$\begin{aligned}
K'_0(s,t) &= 1, \text{for all } s,t,\\
K'_i(s,t) &= 0, \text{if } \min(|s|,|t|) < i,\\
K_i(s,t) &= 0, \text{if } \min(|s|,|t|) < i,\\
K'_i(sx,t) &= \lambda K'_i(s,t)\\
&\quad + \textstyle\sum_{j:\, t_j = x} K'_{i-1}(s, t[1:j-1])\lambda^{|t|-j+2},\\
&\qquad\qquad i = 1,...,n-1,\\
K_n(s,t) &= K_n(s,t)\\
&\quad + \textstyle\sum_{j:\, t_j = x} K'_{n-1}(s, t[1:j-1])\lambda^2.
\end{aligned}$$

문자열 커널은 일반적으로 문서의 길이가 길어질수록 값이 커진다. 이에 따른 치우침을 해결하기 위하여 다음과 같이 정규화한다.

$$\tilde{K}(s,t) = \frac{K(s,t)}{\sqrt{K(s,s)K(t,t)}}$$

4. 실험 및 결과

4.1 실험 데이터

실험을 위해서 네이버 영화 사이트에서 영화 2,082편에 대한 169만개의 40자 리뷰를 데이터로 수집하였다. 수집된 데이터는 인터넷 사용자들이 영화를 1~10점으로 평가한 리뷰로, 실험에서는 1~5점의 리뷰는 부정으로, 6~10점의 리뷰는 긍정으로 선택하였다.

총 169만개의 데이터 중, 실제 실험에 사용할 표본 추출을 위하여 전체 데이터를 형태소 분석기를 이용하여 리뷰에 포함 되어 있는 명사의 비율이 0~20%, 20~40%, 40~60%, 60~80%, 80~100%인 데이터로 분류하였다. 그리고 각 5개의 구간에서 1~10점 리뷰를 각각 800개씩 랜덤하게 추출하였다. 이 때, 문장에 영어,

일어, 한자가 포함되어 있는 리뷰는 추출 대상에서 제외하였다. 즉, 각 구간마다 8,000개의 리뷰가 존재한다. 실험에 사용된 모든 리뷰에 대해 전처리로 구두점 특수기호를 삭제하였다.

4.2 Baseline

제안한 모델의 성능을 증명하기 위하여 baseline으로 자연어 처리 분야에서 일반적으로 사용되는 BOW 모델을 사용하였다. 자질 벡터를 생성하기 위하여 형태소 분석을 통해 내용어(content words)를 추출하고, 추출된 내용어에서 체언, 용언, 기타(관형사, 부사, 감탄사)로 분석된 어구만을 자질로 선택하였다. 일반적으로 자연어 처리 문제에서 BOW 모델에 TF-IDF를 가중치로 사용한다. 하지만 기존 연구에 따르면[2], 감정 문서 분류에서는 TF-IDF보다 단어의 존재 유무를 자질 값으로 사용하는 것이 더 나은 성능을 보인다고 알려져 있다. 이에 본 실험에서는 TF-IDF를 이용하지 않고, 자질 값으로 문서에 포함되어 있는 단어일 경우는 1, 아니면 0을 설정하였다.

아래의 모든 실험에서는 10-fold cross validation을 수행하여 실험 결과가 안정적으로 나왔음을 확보하였다

4.3 실험 결과

표 1은 각 구간에서 존재하는 등록된 명사의 개수를 보여준다. 실험에서 사용한 형태소 분석기는 사전에 등록되어 있지 않은 어구를 미등록 명사로 분류한다 리뷰 문장에 포함된 명사 비율이 높아질수록 등록된 명사의 비율이 낮아진다. 특히 80~100% 구간에서는 미등록된 명사의 개수가 등록된 명사로 분류된 개수보다 4.8배정도 많았다. 이것은 인터넷 리뷰에 띄어쓰기 오류가 빈번하고 신조어를 포함한 새로운 단어가 널리 사용되고 있어 형태소 분석기로는 올바른 형태소 분석이 이루어질 수 없음을 보여준다.

표 1. 각 구간별 등록/미등록 명사의 분포

	등록된 명사		미등록된 명사	
	개수	비율(%)	개수	비율(%)
0 ~ 20%	17,472	49.78	17,629	50.22
20 ~ 40%	17,886	43.13	23,586	56.87
40 ~ 60%	12,701	33.04	25,743	67.96
60 ~ 80%	12,651	32.86	25,844	67.14
80 ~ 100%	4,808	17.23	23,102	82.77

그림 2는 각 문서에 포함된 명사 비율에 따른 성능 변화를 보여준다. BOW 모델은 리뷰에 명사를 포함하는 비율이 높아질수록 성능이 크게 변화하였다. 특히, 문장에 명사가 포함된 비율이 80~100%인 데이터의 경우는 BOW 모델의 성능 하락이 크게 나타났다. 그러나 본 논문에서 제안한 문자열 커널 모델은 명사의 포함 비율이 매우 낮은, 0~20% 구간을 제외하고는 BOW 모델에 비해 훨씬 나은 성능을 보여주었다. 특히 명사의 비율이 높은 데이터에서도 성능 하락이 적어, BOW 모델에 비해 훨씬 안정적인 성능을 보여주었다.

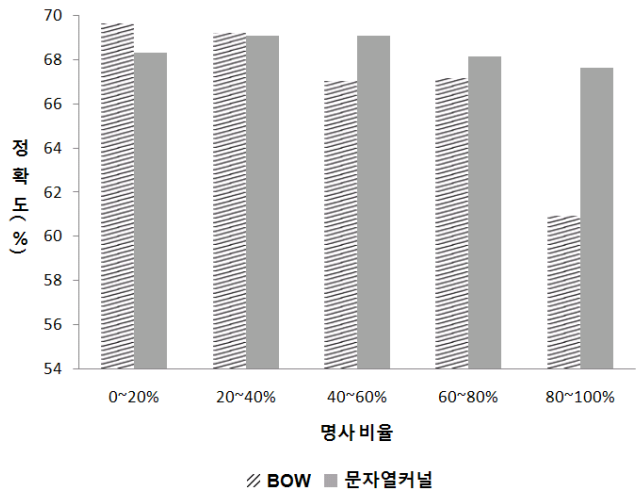

그림 2. 리뷰에 포함된 명사 비율에 따른 BOW 와 문자열 커널 성능 비교

표 2는 전체 169만 리뷰 데이터에 명사 비율이 높은 문서가 실제로 많이 존재하고 있음을 보여준다. 이 표에 따르면, 문자열 커널 모델에서 더 나은 성능을 보여주고 있는 데이터가 전체 데이터의 약 80%에 이름을 알 수 있다. 그러므로 문자열 커널이 0~20% 명사 비율인 구간 데이터에서는 BOW 모델에 조금 낮은 성능을 보여주지만, 전체적인 데이터 구간에서는 더 나은 성능을 보여준다.

표 2. 명사 비율에 따른 데이터의 비율

	개수	비율(%)
0 ~ 20%	164,731	9.73
20 ~ 40%	765,525	45.23
40 ~ 60%	506,223	29.91
60 ~ 80%	134,250	7.93
80 ~ 100%	121,968	7.21

5. 결론 및 향후 과제

본 논문에서는 감정 분석을 위한 새로운 방법을 제안

하였다. 기존의 감정 분석 연구들은 여러 가지 문제점으로 인해 인터넷 문서의 감정 분석에 적용하기가 적합하지 않았다. 이를 해결하기 위하여, 제안한 모델은 추가적인 외부 리소스를 이용하지 않고, 문서의 부분 문자열을 고려한 문자열 커널을 이용하였다. 실험 결과에서 BOW 모델은 리뷰에 포함된 명사 비율이 높을수록 큰 성능 저하를 보였다. 하지만, 제안한 모델은 BOW 모델에 비해 훨씬 안정적인 성능을 보였다.

향후 연구로는 관점이 다른 두 모델을 결합하여 더 나은 성능을 얻을 수 있는 모델을 제안할 것이다. 본 논문에서 실험을 위해 사용된 BOW 모델과 문자열 커널 모델의 오류 데이터를 분석 해 본 결과, 두 모델이 데이터를 서로 다른 관점에서 분류하고 있음을 발견하였다. 그러므로 이 두 모델을 결합하여 서로의 오류를 보완해 줌으로써 더 나은 성능을 보일 수 있는 모델에 대한 연구가 필요할 것으로 보인다.

감사의 글

본 논문은 한국산업기술평가관리원의 정보 통신 선도기반 기술개발사업(A1100-0601-0102)의 연구 결과로 수행 되었습니다.

참 고 문 헌

[1] http://sentiwordnet.isti.cnr.it

[2] 황재원, 고영중, 감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템, 정보과학회논문지, 제 14권, 제3호, pp. 336-340, 2008.

[3] 남상협, 나승훈, 이예하, 이용훈, 김준기, 이종혁, 의견 어구 추출을 위한 생성 모델과 분류 모델을 결합한 부분 지도 학습 방법, 한국컴퓨터종합학술대회 논문집, Vol.35, No.1(C), pp. 268-273, 2008.

[4] 김묘실, 강승식, SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현, 한글 및 한국어 정보처리 학술대회, pp.285-289, 2006.

[5] Huma Lodhi, Craig Saunders, Jojn Shawe-Taylor, Nello Cristianini, and Chris Watkins, "Text Classification using String Kernels," *Journal of Machine Learning Research*, pp. 419-444, 2002

[6] Minqing Hu, and Bing Liu, "Mining and Summarizing Customer Reviews," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177, 2004

[7] Peter D. Turney, "Thumbs up or thumbs down? Semantic Orientation applied to unsupervised classification of reviews," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL'02)*, pp. 417-424, 2002.

[8] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol 10, pp. 79-86, 2002.

[9] Tony Mullen and Nigel Collier., "Sentiment analysis using support vector machines with diverse information sources," *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pp. 412-418, 2004.

[10] Paula Chesley, Bruce Vincent, Li Xu, Rohini K. Srihari, "Using Verbs and Adjectives to Automatically Classify Blog Sentiment," *Proceedings of American Association for Artificial Intelligence - Spring Symposium Series Technical Reports*, pp. 27-30, 2006.

[11] 김명관, 박영택, 감정요소를 사용한 정보검색에 관한 연구, 정보처리학회 논문지 B, 제10-B권, 제6호, pp. 579-586, 2003.

[12] 백선경, 김판구, 신문기사의 감정추출 방법에 관한 연구, 한국컴퓨터종합학술대회 논문집, Vol.32, No.1(B), pp. 562-564, 2005.

[13] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," *Cambridge University Press*, 2000.