Chapter4. Linguistic applications of classification

Natural Language Natural Language Processing (Course materials for Georgia Tech CS 4650 and 7650, "Natural Language")[https://github.com/jacobeisenstein/gt-nlp-class]

Chapter4. 목차

4.1 Sentiment and opinion analysis

- **텍스트 분류의 보편적인 적용**
  - 제품 리뷰 및 소셜 미디어 게시물과 같은 문서의 정서 또는 의견의 극성을 자동으로 결정하는 것
  - 예를 들어, 마케팅 담당자는 사람들이 광고, 서비스 및 제품에 어떻게 반응 하는지를 알고 싶어합니다 (Hu and Liu, 2004).
  - 사회 과학자들은 날씨 (Hannak et al., 2012)와 같은 현상에 의해 감정이 어떻게 영향을 받는지와 사회적 네트워크를 통해 의견과 감정이 모두 어떻게 퍼지는 지에 관심이있다 (Coviello et al., 2014; Miller et al., 2011).
  - 디지털 인문학 분야의 문학 학자들은 소설을 통해 감정의 흐름을 통해 음모 구조를 추적합니다 (Jockers, 2015).

- **감정 분석은 신뢰할 수있는 레이블을 얻을 수 있다고 가정할 때 문서 분류의 직접적인 적용으로 구성**
  - 가장 단순한 경우 정서 분석은 POSITIVE, NEGATIVE 및 NEUTRAL의 정서 (2개 또는 3개)
  - 행복한 이모티콘이 포함 된 트윗은 긍정적인 것으로 표시 될 수 있으며 슬픈 이모티콘은 부정적으로 표시 될 수 있습니다.
  - 4 개 이상의 별을 사용한 리뷰는 양성으로 표시 될 수 있으며, 2 개 이하의 별은 음수로 표시 될 수 있습니다.
  - 주어진 법안에 찬성표를 던지는 정치인들의 진술은 (그 법안에 대해) 긍정적 인 것으로 표시됩니다. 법안에 반대하는 정치인들의 진술은 부정적으로 표시됩니다.

- **bag-of-words 모델은 문서 수준에서 정서 분석에 적합함.**

4.1 Sentiment and opinion analysis

- Lexicon-based classification은 단일 문장 또는 소셜 미디어 게시물과 같은 짧은 문서에는 효과적이지 않음
- **Example**

(4.1) That's not bad for the first day.

(4.2) This is not the worst thing that can happen.

(4.3) It would be nice if you acted like you understood.

(4.4) There is no reason at all to believe that the polluters are suddenly going to become reasonable. (Wilson et al., 2005)

(4.5) This film should be brilliant. The actors are first grade. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great plot, **however**, the film is a failure. (Pang et al., 2002)

A minimal solution is to move from a bag-of-words model to a bag-of-**bigrams** model, where each base feature is a pair of adjacent words, e.g.,

$$(that's, not), (not, bad), (bad, for), \ldots \quad [4.1]$$

4.1.1 Related problems

- **Subjectivity detection**
  - ✓ 주관적인 견해를 표현하는 텍스트 부분과 추측과 가설과 같은 기타 비 사실적 내용을 식별해야 함(Riloff and Wiebe, 2003).
  - ✓ 이것은 각 문장을 별도의 문서로 취급 한 다음 bag-of-words를 적용하여 수행 할 수 있습니다
  - ✓ 근처의 문장이 동일한 레이블을 갖도록 하는 그래프 기반 알고리즘을 사용하여 bag-of-words 모 보강

- **Stance classification**
  - ✓ 각 참가자는 채식 생활방식을 채택하거나 무료 대학 교육을 요구하는 등의 제안을 찬성하거나 반대하는 등의 측면을 지지합니다.
  - ✓ 논증의 텍스트에서 저자의 지위를 확인하는 것. 경우에 따라 각 위치별로 사용할 수있는 훈련 데이터가 있으므로 표준 문서 분류 기술을 사용할 수 있음.
  - ✓ 가장 어려운 경우에는 입장에 대한 분류된 데이터가 없으므로 유일한 위치는 동일한 직책을 지지하는 그룹 문서입니다 (제 5 장에서 논의된 자율학습의 한 형태)

- **Targeted sentiment analysis**

  (4.6) The vodka was good, but the meat was rotten.

  (4.7) Go to Heaven for the climate, Hell for the company. –*Mark Twain*
  - ✓ 이러한 진술은 혼합된 전반적인 감정을 나타내며 일부 엔티티 (예 : 보드카)에 대해서는 양성이고 다른 그룹 (예 : 고기)에 대해서는 음수, 특정 실체에 대한 작가의 감정을 확인하고자 함 (Jiang et al., 2011). 이를 위해서는 텍스트의 엔티티를 식별하여 특정 감정 단어와 연결해야 함

- **Aspect-based opinion mining**
  - Price and service 관점으로 identify the sentiment of the author of a review

- **Emotion classification**
  - 심리학자는 일반적으로 감정을 보다 다각적인 것으로 간주, 행복, 놀라움, 두려움, 슬픔, 분노 및 멸시의 6 가지 기본 감정이 있으며 인간 문화 전반에 걸쳐 보편적이라고 주장
  - 이야기 형식 (예 : 농담, 민화)을 포착하는 기능 및 이야기의 각 문장의 위치를 반영하는 구조적 특징을 사용하여 단어중심 기능 (인간 주석 자조차도 서로 의견이 맞지않음, 60~70% 정확도)

4.1.2 Alternative approaches to sentiment analysis

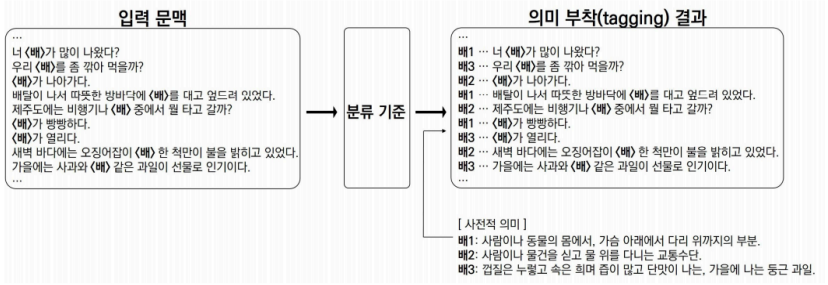
- **Regression**
  - ✓ 수치 척도 (Pang and Lee, 2005)를 결정하는 것, 제곱 오차를 최소화하는 가중치 θ를 식별하는 것
  - ✓ 가중치가 페널티 λ || θ || 2를 사용하여 정규화된다면 그것은 릿지 회귀

- **Ordinal ranking**
  - ✓ 서수 순위 문제에서 레이블은 순서가 있지만 Discrete, (Ex) 제품 리뷰는 종종 1 - 5의 눈금으로 표시되며 점수는 A - F의 눈금
  - ✓ 점수 θ · x를 "rank"로 이산화하여 해결할 수 있음, 퍼셉트론과 같은 알고리즘을 사용하여 가중치와 경계를 동시에 학습

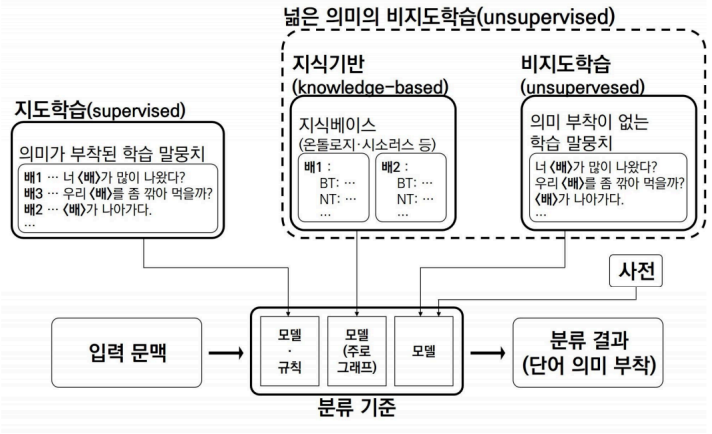
- **Lexicon-based classification**
  - ✓ 손으로 만들어진 기능 가중치가 여전히 널리 사용되고 있는 유일한 NLP 작업 중 하나
  - ✓ 어휘 기반 분류 (Taboada et al., 2011)에서 사용자는 각 레이블에 대한 단어 목록을 작성한 다음 각 목록의 단어 수에 따라 각 문서를 분류
  - ✓ 초기 어휘집은 General Inquirer (Stone, 1966)이었다. 오늘날 인기있는 정서 어휘집에는 sentiwordnet (Esuli and Sebastiani, 2006)
  - ✓ 세분화된 분석을 위해 언어 문의 및 단어 수 (LIWC)는 일련의 어휘를 제공 (Tausczik and Pennebaker, 2010)
  - ✓ MPQA 사전은 8221 용어의 극성 (양수 또는 음수)과 해당 용어가 강하게 또는 약하게 주관적이다 (Wiebe et al., 2005).
  - ✓ 정서 어휘집의 포괄적 인 비교는 Ribeiro et al. (2016). 초기 종자 사전이 주어지면 종자 세트에서 단어와 자주 발생하는 단어를 찾아서 자동으로 사전을 확장 할 수 있습니다 (Hatzivassiloglou and McKeown, 1997; Qiu et al., 2011).

4.2 Word sense disambiguation

- 단어 의미 중의성 해소 또는 어의 중의성 해소는 같은 형태의 단어라도 문 맥에 따라 다른 의미로 사용되었을 수 있다는 사실을 전제로, 주어진 문맥(주로 문장이나 문단 단위)에서 특정 단어가 어떤 사전적 의미로 쓰였는지 구분 해내는 연구 분야



<그림 1> 문맥 분류의 관점에서 본 단어 의미 중의성 해소 과정



<그림 2> 단어 의미 중의성 해소 연구의 방법론에 따른 분류

(4.8) Iraqi head seeks arms (이라크 수뇌부, 무장을 추구하다. / 이라크인의 머리, 팔을 찾다)
(4.9) Prostitutes appeal to Pope (매춘부는 교황에게 호소한다. / 매춘부는 교향에게 매력적이다)
(4.10) Drunk gets nine years in violin case (주정뱅이, 바이올린 사건으로 9개월 선고 받다.
/ 주정뱅이, 바이올린 케이스에 9개월 동안 갇히다.)

각 단어 토큰에 대한 올바른 의미를 식별하는 것입니다. 품사 모호성 (예를 들어, 명사 대 동사)은 일반적으로 초기 단계에서 해결되는 다른 문제로 고려된다. 언어적 관점에서 보면, 감각은 단어의 속성이 아니라 표제어로, 굴절된 단어 집합을 대표하는 표준 형태입니다.
예를 들어, arm / N은 굴절된 형태의 arms / N을 포함하는 보조 정리. / N은 우리가 명사를 가리키는 것을 나타내며 동음이의 arm / V는 굴절 된 동사를 포함하는 또 다른 보조 정리가 아닙니다 / V, arms / V, armed / V, arming / V). 따라서 단어 감별은 먼저 각 토큰에 대한 정확한 품사와 표제어를 식별 한 다음 해당 토큰에 연결된 인벤토리에서 올바른 감각을 선택해야 합니다

4.2 How many word senses?

Words sometimes have many more than two senses, as exemplified by the word *serve*:

- [FUNCTION]: *The tree stump served as a table*
- [CONTRIBUTE TO]: *His evasive replies only served to heighten suspicion*
- [PROVIDE]: *We serve only the rawest fish*
- [ENLIST]: *She served in an elite combat unit*
- [JAIL]: *He served six years for a crime he didn't commit*
- [LEGAL]: *They were served with subpoenas*[4]

의미의 차이점은 영어에 대한 어휘 의미 데이터베이스 인 WordNet (http://wordnet.princeton.edu)에 주석으로 표시됨. WordNet은 약 10 만 개의 synset으로 구성되어 있는데, 이는 동의어인 보조 정리 (또는 구) 그룹입니다. 예제 synset은 다음과 같습니다. {chump1, fool2, sucker1, mark9}

WordNet은 단어 감각 불균형 문제의 범위를 정의하며, 더 일반적으로 영어의 어휘 의미론적 지식을 형식화. (WordNet은 수십 가지 다른 언어로 다양한 수준으로 작성되었습니다.)
* 기타 어휘 의미 관계 동의어 외에도 WordNet은 다음을 포함하여 많은 다른 어휘 의미 관계를 설명합니다.
• antonymy : x는 y의 반대를 의미합니다.
• hyponymy : x는 y의 특별한 경우입니다
• meronymy : x는 y의 일부입니다 (예 : WHEEL-BICYCLE).

4.2.2 Word sense disambiguation as classification

How can we tell living *plants* from manufacturing *plants*? The context is often critical:

(4.11) Town officials are hoping to attract new manufacturing plants through weakened environmental regulations.

(4.12) The endangered plants play an important role in the local ecosystem.

It is possible to build a feature vector using the bag-of-words representation, by treating each context as a pseudo-document. The feature function is then,

$$f((\textit{plant}, \textit{The endangered plants play an} \ldots), y) = \\ \{(\textit{the}, y) : 1, (\textit{endangered}, y) : 1, (\textit{play}, y) : 1, (\textit{an}, y) : 1, \ldots\}$$

As in document classification, many of these features are irrelevant, but a few are very strong predictors. In this example, the context word *endangered* is a strong signal that the intended sense is biology rather than manufacturing. We would therefore expect a learning algorithm to assign high weight to $(\textit{endangered}, \text{BIOLOGY})$, and low weight to $(\textit{endangered}, \text{MANUFACTURING})$.[5]

It may also be helpful to go beyond the bag-of-words: for example, one might encode the position of each context word with respect to the target, e.g.,

$$f((\textit{bank}, \textit{I went to the bank to deposit my paycheck}), y) = \\ \{(i-3, \textit{went}, y) : 1, (i+2, \textit{deposit}, y) : 1, (i+4, \textit{paycheck}, y) : 1\}$$

These are called **collocation features**, and they give more information about the specific role played by each context word. This idea can be taken further by incorporating additional syntactic information about the grammatical role played by each context feature, such as the **dependency path** (see chapter 11).

4.3 Design decisions for text classification

4.3.1 What is a word?

4.3.1.1 Tokenization

bag-of-words 벡터를 구성하기위한 첫 번째 하위 작업은 토큰화. 텍스트를 문자 시퀀스에서 단어 토큰 시퀀스로 변환함. 간단한 접근법은 문자의 부분 집합을 공백으로 정의한 다음 텍스트를이 토큰으로 분할하는 것

Whitespace	Isn't	Ahab,	Ahab?	;)				
Treebank	Is	n't	Ahab	,	Ahab	?	;	)
Tweet	Isn't	Ahab	,	Ahab	?	;)		
TokTok (Dehdari, 2014)	Isn	'	t	Ahab	,	Ahab	?	;)

Figure 4.1: The output of four `nltk` tokenizers, applied to the string *Isn't Ahab, Ahab? ;)*

공백 기반 토큰화는 이상적이지 않음. *prize-winning. half-asleep 단어는* 하이픈을 넣은 구문으로 분리. 단어를 쉼표와 마침표 다음에 구분하는 것이 좋습니다. 동시에 미국과 박사 같은 약어를 분리하지 않는 것이 좋음.
토큰화는 대개 정규 표현식을 사용하여 수행. 예를 들어, nltk 패키지에는 많은 토큰화 프로그램이 포함되어 있음 토큰 화는 언어 별 문제이며 각 언어마다 고유한 문제가 있음.

- 예를 들어, 중국어에는 단어 사이에 공백이나 단어 경계의 다른 일관된 직교 표식이 포함되지 않습니다. "욕심 많은"접근법은 사전 정의된 어휘집에 있는 문자 하위 문자열에 대한 입력을 스캔하는 것입니다. 그러나, Xue et al. (2003)은 많은 문자 시퀀스가 여러 방법으로 분할 될 수 있기 때문에 이것이 모호 할 수 있다고 지적합니다. 대신 그는 각 중국어 문자 또는 한자가 단어 경계인지 여부를 결정하기 위해 분류자를 훈련시킵니다. 단어 세분화를위한 보다 진보 된 서열 분류 방법은 8.4에서 논의된다.
- 유사한 문제는 독일어와 같은 알파벳 스크립트를 사용하는 언어에서도 발생할 수 있습니다. 예를 들어, Freundschaftsbezeigungen (우정의 시위) 및 Dilettantenaufdringlichkeiten (딜레간티의 수입품)과 같은 예를 산출하는 복합 명사의 공백을 포함하지 않습니다.

트웨인 (Twain, 1997)은 "이 단어는 단어가 아니며 영문자 행렬"이라고 주장하고있다. 소셜 미디어는 #TrueLoveInFourWords와 같은 해시 태그를 사용하여 분석을 위해 분해가 필요한 영어 및 다른 언어에서도 비슷한 문제를 제기한다 (Brun and Roux, 2014).

4.3 Design decisions for text classification

4.3.1 What is a word?

4.3.1.2 Normalization

텍스트를 토큰으로 분할한 후, 다음 질문은 토큰이 정말로 구별되는 것입니다. *great*, *Great*, and GREAT 을 구분할 필요가 있습니까?
대소문자 구별을 완전히 제거하면 더 작은 어휘가 생기므로 더 작은 특징 벡터가 생깁니다. 그러나 경우에 따라 대소문자가 구별 될 수 있습니다. 예를 들어 *apple* 은 맛있는 파이를, *Apple* 은 company that specializes in proprietary dongles and power adapters.

Original	The	Williams	sisters	are	leaving	this	tennis	centre
Porter stemmer	the	william	sister	are	leav	thi	tenni	centr
Lancaster stemmer	the	william	sist	ar	leav	thi	ten	cent
WordNet lemmatizer	The	Williams	**sister**	are	leaving	this	tennis	centre

Figure 4.2: Sample outputs of the Porter (1980) and Lancaster (Paice, 1990) stemmers, and the `WordNet` lemmatizer

로마 스크립트의 경우 유니코드 문자열 라이브러리를 사용하여 대 / 소문자 변환을 수행 할 수 있음
대다수 스크립트는 대소문자 구분하지 않음, 대 / 소문자 변환은 정규화의 한 유형, 다른 표준화에는 숫자의 표준화 (예 : 1,000-1000) 및 날짜 (예 : 2015 년 8 월 11 일 ~ 2015/11/08)가 포함됩니다. 소셜 미디어에는 cooooool과 같이 표현 길이를 길게하는 등의 정규화 될 수있는 추가 정형 현상이 있습니다
극단적인 정규화 형식은 영어로 -ed 및 -s 접미어와 같은 굴절 식 접미어를 제거하는 것

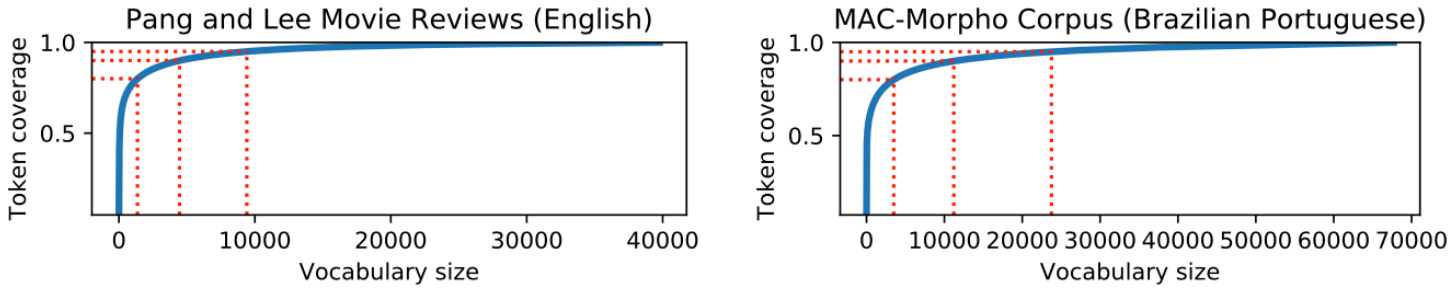
Stemmer : 일련의 정규 표현식을 적용하여 접미사를 제거하는 프로그램

4.3 Design decisions for text classification

4.3.1 What is a word?

4.3.1.2 Normalization

Lemmatizers : 주어진 wordform의 근원적인 보조 정리를 식별하는 시스템, 그림 4.2에서 형태소 분석자의 과도기 론적 오류를 피하고 *geese→goose* 와 같은 좀 더 복잡한 변형을 처리함



(a) Movie review data in English (b) News articles in Brazilian Portuguese

Figure 4.3: Tradeoff between token coverage (y-axis) and vocabulary size, on the `nltk` movie review dataset, after sorting the vocabulary by decreasing frequency. The red dashed lines indicate 80%, 90%, and 95% coverage.

정규화의 가치는 데이터 및 작업에 따라 다릅니다. 정규화는 기능 공간의 크기를 줄여 일반화에 도움이 됨. 그러나 언어적으로 의미있는 구분을 병합 할 위험은 항상 있습니다. 감독 된 기계 학습에서 정규화 및 다듬기는 정확한 정규화에 필요한 언어 별 엔지니어링을 피하면서 학습자가 희귀 한 기능에 지나치게 적응하지 못하도록 표준화와 유사한 역할을 수행 할 수 있습니다.
컨텐츠 기반 정보 검색 (Manning et al., 2008)과 주제 모델링 (Blei et al., 2003)과 같은 감독되지 않은 시나리오에서 정규화가 더 중요합니다.

4.3 Design decisions for text classification

4.3.2 How many words?

특징 벡터의 크기를 제한하면 결과모델의 메모리 공간이 줄어들고 예측 속도가 빨라짐. 정규화는 이러한 역할을하는 데 도움이 될 수 있지만 더 직접적인 접근은 어휘를 데이터 집합에서 N 개의 가장 빈번한 단어로 제한하는 것

- 예를 들어, nltk (원래 Pang et al., 2002)에서 제공 한 영화 리뷰 데이터 세트에는 39,768 개의 단어 유형과 1.58M 토큰이 있습니다. 그림 4.3a에서와 같이 가장 빈번한 4000 개의 단어 유형이 모든 토큰의 90 %를 차지하며 모델 크기의 크기 감소를 제공합니다. 예를 들어 포르투갈어 Mac-Morpho 브라질어 (Alu ısio 외., 2003)에서 90 % 적용 범위를 달성하려면 10000 개 이상의 단어 유형이 필요합니다 (그림 4.3b). 이것은 포르투갈어의 형태 학적 복잡성을 반영하며 영어보다 많은 굴절식 접미사를 포함합니다.

- 드문 단어를 제거하는 것이 분류 성능에 항상 유리한 것은 아닙니다. 예를 들어, 일반적으로 드문 이름은 뉴스 기사의 주제를 구별하는 데 큰 역할을 합니다.
- 피쳐 공간의 크기를 줄이는 또 다른 방법은,,, 및와 같은 **stopwords를** 제거하는 것 (주제, 감정 또는 자세를 표현할 때 거의 역할을하지 않는 것처럼 보일 수 있습니다.)
- 그러나 코퍼스 언어 학자와 사회 심리학자들은 겉으로보기에는 중요하지 않은 단어가 놀라운 통찰력을 제공 할 수 있음을 보여주었습니다
- 정규화와 마찬가지로 용어 기반 문서 검색과 같은 관리되지 않는 문제의 경우 스톱워드 필터링이 더 중요합니다.

- 모델 크기를 제어하기위한 또 다른 대안은 피쳐 해싱 (feature hashing)이다 (Weinberger et al., 2009).
  - 각 기능에는 해시 기능을 사용하여 색인이 지정됨. 충돌을 허용하는 해시 함수가 선택되면 (일반적으로 해시 출력을 일부 정수로 가져옴으로써) 여러 모델이 단일 가중치를 공유하기 때문에 모델을 임의로 작게 만들 수 있습니다.

4.3 Design decisions for text classification

4.3.3 Count or binary?

마지막으로,

우리는 우리의 특징 벡터가 각 단어의 수 또는 그 존재를 포함하기를 원할지를 고려할 수있다.

선형 분류의 미묘한 한계에 도달 : 하나보다 두 실패가 더 나빠지지만, 실제로는 두 배나 나쁜 것인가? 이 직감에 의해 동기 부여 된 Pang et al. (2002)는 특징 벡터에 존재 또는 부재의 이진 지시자를 사용한다 : $f_j(x, y) \in \{0, 1\}$.

그들은 이 바이너리 벡터에 대해 훈련 된 분류자 (classifier)가 단어 수에 따라 특징 벡터를 능가하는 경향이 있음을 발견했다. 한 가지 설명은 단어가 덩어리로 나타나는 경향이 있다는 것입니다. 한 단어가 문서에 한 번 나타나면 다시 나타납니다 (Church, 2000).

이러한 후속 출현은 반복을 향한 이러한 경향에 기인 할 수 있으며, 따라서 문서의 등급 라벨에 대한 추가 정보를 거의 제공하지 않는다.

4.3 Design decisions for text classification

4.4 Evaluating classifiers

감독된 기계학습 프로그램에서 held-out test set 를 지정하는 것이 중요.

- 이 데이터는 단일 분류 자의 전체 정확도를 평가하는 한 가지 목적으로만 사용해야합니다. 이 데이터를 두 번 이상 사용하면 분류 기준 이 이 데이터에 맞게 사용자 정의되고 미래의 보이지 않는 데이터뿐만 아니라 수행되지 않기 때문에 예상 정확도가 지나치게 낙관적으로 나타날 수 있습니다.
- 하이퍼 파라미터를 설정하거나 기능 선택을 수행하는 것이 일반적으로 필요하므로 목적을 위해 튜닝 또는 개발 세트를 구성해야 할 수도 있습니다.
- 분류 기준의 성능을 평가하는 데는 여러 가지 방법이 있습니다. 가장 간단한 방법은 정확도입니다. 정확한 예측 수를 총 인스턴스 수로 나눈 값입니다.

다른 측정 항목이 필요한 이유는 무엇입니까? 주된 이유는 계급 불균형입니다.
전자 건강 기록 (EHR)이 드문 질병의 증상을 설명하는지 여부를 감지하기위한 분류자를 작성한다고 가정합니다.

- 이 질병은 데이터 세트의 모든 문서 중 단 1 % 만 나타납니다. 모든 문서에 대해 y = 음수를 보고하는 분류기는 99 %의 정확도를 얻지 만 실제로 쓸모가 없습니다. 분배가 비뚤어 질 때조차도 클래스를 구별하는 분류 자의 능력을 탐지 할 수있는 메트릭이 필요합니다.
- 한 가지 해결책은 각 가능한 레이블이 동등하게 표현되는 균형 테스트 세트를 만드는 것입니다. 그러나 EHR 예에서 이것은 원래 데이터 세트의 98 %를 버리는 것을 의미합니다! 또한 탐지 임계 값 자체가 설계 고려 사항 일 수 있음.
- 건강 관련 응용 프로그램에서 매우 민감한 분류자를 선호 할 수 있습니다. y (i) = POSITIVE라는 작은 기회가 있을 경우 긍정적 예측을 반환합니다. 다른 응용 프로그램에서는 긍정적 인 결과가 값 비싼 행동을 유발할 수 있으므로 절대적으로 확실한 경우에만 긍정적인 예측을 하는 분류 기능을 선호합니다. 추가 측정 항목이 필요합니다.

4.3 Design decisions for text classification

4.4 Evaluating classifiers

Confunsion Matrix (https://en.wikipedia.org/wiki/Confusion_matrix).

Evaluating multi-class classification

- 다중 클래스 분류 평가 recall, precision 및 F -MEASURE는 특정 레이블 k에 대해 정의됩니다.
- 관심있는 레이블이 여러 개인 경우 (예 : 단어 감별 또는 감정 분류) 각 클래스에 F-MEASURE를 결합해야합니다.
- **Macro** F -MEASURE는 여러 클래스에 걸친 평균 F -MEASURE입니다. 불균형 클래스 분포가 있는 다중 클래스 문제에서 **Macro F -MEASURE**는 분류자가 클래스를 얼마나 잘 인식하는지 균형있게 측정 한 것입니다.
- **Micro** F -MEASURE에서는 각 클래스에 대해 false positives, false positive 및 false negative을 계산한 다음 이를 추가하여 단일 리콜, 정밀도 및 F -MEASURE를 계산합니다. 이 측정 항목은 균형적 입니다.
- 클래스가 아닌 인스턴스 전체에 적용되므로 각 클래스에 균등하게 가중치를 부여하는 **Macro** F -MEASURE와 달리 각 클래스를 빈도에 비례하여 가중치를 부여합니다.

- 설명 : http://operatingsystems.tistory.com/entry/Data-Mining-Macro-average-and-micro-average

4.3 Design decisions for text classification

4.4.2 Threshold-free metrics

이진 분류 문제에서는 점수 함수의 출력에 상수 "**임계 값**"을 추가하여 리콜과 정밀도를 절충하는 것이 가능합니다. 커브를 추적 할 수 있습니다. 각 점은 단일 임계 값에서 성능을 나타냅니다. ROC 곡선은 곡선의 면적 (AUC)을 적분하여 단일 숫자로 요약 할 수 있습니다.

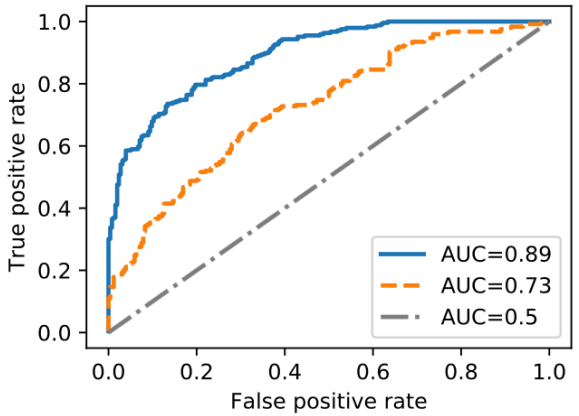


Figure 4.4: ROC curves for three classifiers of varying discriminative power, measured by AUC (area under the curve)

AUC는 랜덤하게 선택된 positive의 예가 무작위로 선택된 네거티브 예보다 분류기에 의해 높은 점수가 할당 될 확률로 해석 될 수 있음. F MEEAURE에 비해 AUC의 한 가지 이점은 0.5의 기준선 비율이 라벨 분포에 의존하지 않는다는 것입니다.

4.3 Design decisions for text classification

4.4.3 Classifier comparison and statistical significance

자연어 처리 연구 및 엔지니어링은 종종 서로 다른 분류 기법을 비교하는 것을 포함합니다.
- 모델 비교는 로지스틱 회귀 대 평균화 된 퍼셉트론 또는 L2 정규화 대 알고리즘과 같은 알고리즘 간의 비교.
- 피처 비교는 bag-of-word 대 position bag-of-word (§ 4.2.2 참조)와 같은 기능 세트 사이의 비교.

Ablation testing는 피쳐 그룹과 같은 분류기의 다양한 측면을 체계적으로 제거 (제거)하고 절제 된 분류자가 전체 모델만큼 우수하다는 귀무 가설을 테스트하는 것을 포함합니다.
가설 테스트의 주요 목표는 두 통계의 차이 (예 : 두 분류 기준의 정확성)가 우연히 발생할 가능성이 있는지를 결정하는 것.

4.3 Design decisions for text classification

4.4.3.1 The binomial test

정확성의 차이에 대한 통계적 유의성은 이항 테스트와 같은 고전적 테스트를 사용하여 평가할 수 있습니다.

분류기 c1과 c2가 바이너리 레이블이 있는 테스트 세트에서 N 개의 인스턴스에 대해 동의하지 않으며 해당 인스턴스의 k에서 c1이 정확하다고 가정합니다 .

분류기가 똑같이 정확하다는 귀무가설 하에서, 우리는 k / N이 대략 1/2과 같을 것이고, N이 증가함에 따라, k / N은 이 기대치에 점차 근접해야 할 것입니다. 이진확률 변수의 개수에 대한 확률인 이항 분포에 의해 포착됩니다. k가 이항 분포로부터 이끌려 나오도록 매개 변수 N이 무작위 "드로우"의 수를 나타내고, θ가 각 무승부에서 "성공"확률을 나타내는 k ~ Binom (θ, N)이라고 쓴다.

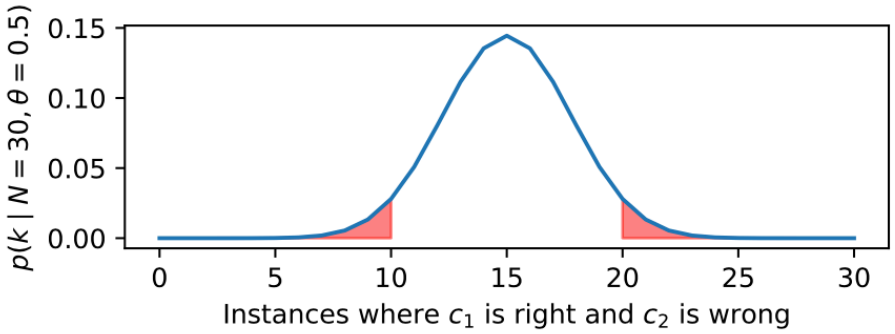


Figure 4.5: Probability mass function for the binomial distribution. The pink highlighted areas represent the cumulative probability for a significance test on an observation of $k = 10$ and $N = 30$.

4.3 Design decisions for text classification

4.4.3.2 *Randomized testing

이항 테스트는 정확도에 적합하지만 F-MEASURE와 같은 더 복잡한 측정 기준에는 적합하지 않습니다. 임의의 메트릭에 대한 통계적 유의성을 계산하기 위해 무작위 추출을 적용 할 수 있습니다. **M 부트스트랩 샘플, 교체로 원본 테스트 세트의 인스턴스를 다시 샘플링합니다.**

각 부트스트랩 샘플 자체는 크기 N의 테스트 세트입니다. 원래 테스트 세트의 일부 인스턴스는 주어진 부트 스트랩 샘플에 나타나지 않지만 다른 인스턴스는 여러 번 나타납니다. 전체적으로 샘플은 원래 테스트 세트와 동일한 분포에서 추출됩니다. 그런 다음 각 부트 스트랩 샘플에서 원하는 평가를 계산할 수 있습니다. 그러면 메트릭 값에 대한 분산이 제공됩니다.

두 개 분류기 c1과 c2의 F-MEASURE를 비교하기 위해 함수 $\delta(\cdot)$를 설정하여 부트스트랩 샘플에서 F-MEASURE의 차이를 계산합니다.

그 차이가 표본의 최소 5 %에서 0보다 작거나 같으면 c2가 적어도 c1만큼 우수하다는 단측 귀무 가설을 기각 할 수 없다 (Berg-Kirkpatrick et al., 2012). 또한 단일 분류 기준의 F-MEASURE와 같은 관심 측정 기준을 중심으로 95 % 신뢰 구간에 관심을 가질 수 있습니다

4.3 Design decisions for text classification

4.4.4 Multiple comparisons

여러 데이터 세트에서 여러 분류 자의 성능을 비교할 때와 같이 여러 가설 테스트를 수행해야 하는 경우가 있습니다.

5 개의 데이터 세트가 있고, 분류 기준의 네 가지 버전을 기존 시스템과 비교하여 총 20 개의 비교를 한다고 가정하십시오.

분류 기준 중 어느 것도 분류 기준보다 우수하지 않더라도 결과에 약간의 우연한 차이가 있을 수 있으며
p = 0.05 = 1/20에서 통계적으로 유의 한 개선 효과를 기대할 수 있습니다. 따라서 다중 비교 결과를 보고 할 때 20 개의 p- 값을 조정해야합니다.

하나의 접근법은 m 개의 테스트를 수행 할 때 p < a의 p 값을 보고하기 위해 α의 임계 값을 요구하는 것입니다. 이것은 **Bonferroni correction** 으로 알려져 있음, (a/m threshold)

또 다른 접근법은 FDR(**false discovery rate)**. Benjamini and Hochberg (1995)는 α에서 거짓 발견의 비율을 제한하는 p 값 보정을 제안한다 :
각 개별 테스트의 p 값을 오름차 순으로 정렬하고, pk ≤ k/m X α가되도록 가장 큰 k와 같은 유의도 임계 값을 설정한다.

4.3 Design decisions for text classification

4.5 Building datasets

때로는 분류자를 만들려면 먼저 자신의 데이터 집합을 만들어야합니다. 여기에는 주석을 달기 위해 일련의 문서 또는 인스턴스를 선택한 다음 주석을 수행하는 작업이 포함됩니다. 데이터 세트의 범위는 응용 프로그램에 의해 결정될 수 있습니다.

전자 건강 기록을 분류하기 위한 시스템을 구축하려는 경우 분류기가 배포 할 때 마주 치게 될 유형의 레코드 모음으로 작업해야합니다. 다른 경우, 목표는 광범위한 문서에서 작동 할 시스템을 구축하는 것입니다.

이 경우 여러 스타일과 장르의 공헌으로 균형 잡힌 자료를 갖는 것이 가장 좋습니다. 예를 들어 브라운 코퍼스는 정부 문서에서 로맨스 소설 (Francis, 1964)까지 다양한 텍스트를 사용하며 Google Web Treebank에는 웹 문서의 5 가지 도메인 ("질문 답변, 전자 메일, 뉴스 그룹, 리뷰 및 블로그")에 대한 표기법이 포함되어 있습니다. (Petrov and McDonald, 2012).

4.3 Design decisions for text classification

4.5.1 Metadata as labels

주석은 어렵고 시간이 많이 걸리고 대부분의 사람들은 오히려 그것을 피할 것입니다. 때때로 기존 메타 데이터를 활용하여 분류기를 훈련하기 위한 레이블을 얻을 수 있습니다. 예를 들어 리뷰에는 분류 등급으로 변환 할 수있는 수치 등급이 수반됩니다(§ 4.1 참조).

마찬가지로 소셜 미디어 사용자의 국적은 프로필 (Dredze et al., 2013) 또는 게시물의 시간대 (Gouws et al., 2011)를 통해 추정할 수 있습니다. 보다 야심 차게, 우리는 정치인 및 주요 정당과의 소셜 네트워크 연결을 기반으로 소셜 미디어 프로필의 정치적 제휴를 분류하려고 시도 할 수 있습니다 (Rao 외., 2010).

수동 주석없이 대형 라벨 데이터 세트를 신속하게 구축 할 수있는 편리함은 매력적입니다. 그러나 이 방법은 레이블이없는 인스턴스 (메타 데이터를 사용할 수 없음)가 레이블이있는 인스턴스와 유사 할 것이라는 가정에 의존합니다.

정치인과의 네트워크를 기반으로 한 소셜 미디어 사용자의 정치적 소속을 표시하는 예를 고려해보십시오. 분류 기준이 이러한 테스트 세트에서 높은 정확도를 얻은 경우 모든 소셜 미디어 사용자의 정치적 소속을 정확히 예측한다고 가정하는 것이 안전합니까? 아마도 그렇지 않습니다. 정치인과 소셜 네트워크의 관계를 맺는 소셜 미디어 사용자는 일반 사용자와 비교할 때 정치적 메타 데이터를 사용할 수없는 메시지의 텍스트에서 정치를 언급 할 가능성이 더 높습니다. 그렇다면 소셜 네트워크 메타 데이터로 구성된 테스트 세트의 정확성은 레이블이없는 데이터에 대한 메소드의 진정한 성능을 지나치게 낙관적으로 보여줍니다.

4.3 Design decisions for text classification

4.5.2 Labeling data

manual annotation 이외의 ground truth labels 을 얻을 수있는 방법이 없습니다, 주석 프로토콜은 다음과 같은 몇 가지 기준을 충족해야합니다. 주석은 관심있는 현상을 포착 할 정도로 표현력이 있어야합니다. 복제 가능, 즉, 동일한 데이터가 주어지면 다른 주석자 또는 주석자 팀이 매우 유사한 주석을 생성합니다. 상대적으로 신속하게 생산 될 수 있도록 확장 가능해야합니다.

1. Determine what the annotations are to include
(문서의 저자의 감정 상태에 대한 주석을 생산하는 것, 기본 이론의 전체 인스턴스화는 규모에 따라 주석을 달기에는 너무 비싸므로 합리적인 근사를 고려해야합니다**)**

2. Optionally, one may design or select a software tool to support the annotation effort.
(기존 범용 주석 도구로는 BRAT , MMAX2**)**

3. Formalize the instructions for the annotation task.
(지시사항이 명시적이지 않은 한, 결과 주석은 주석자의 직관에 달려있다. 다른 연구자에 의해 복제되고 사용 가능하도록 하기 위해서는 명시적인 지시가 중요**)**

4. Perform a pilot annotation of a small subset of data, with multiple annotators for each instance
(주석 설명의 복제 가능성과 확장성을 사전에 평가할 수 있습니다**)**

5. Annotate the data.
(주석 프로토콜 및 지침을 완료 한 후 주요 주석 작업을 시작할 수 있습니다. 모든 인스턴스가 아닌 일부 인스턴스는 여러 주석을 받아야하므로 주석 간 계약을 계산할 수 있습니다**)**

6. Compute and report inter-annotator agreement, and release the data.
(경우에 따라 저작권 또는 개인 정보 보호와 관련하여 원시 텍스트 데이터를 공개 할 수 없습니다. 이 경우 한 가지 해결책은 문서 식별자에 대한 링크가 포함 된 독립 주석을 공개적으로 릴리스하는 것입니다. 문서 자체는 라이센스 계약의 조건에 따라 배포 될 수 있으며, 이는 데이터 사용 방법에 조건을 부과 할 수 있습니다. 데이터를 공개 할 때 발생할 수있는 잠재적 결과를 생각하는 것이 중요함)

4.3 Design decisions for text classification

4.5.2.1 Measuring inter-annotator agreement

주석의 복제 가능성을 측정하기 위해 표준적인 방법은 주석자가 서로 동의하는 정도를 계산하는 것.

Annotators 가 동의하지 않으면 신뢰성 또는 어노테이션 시스템 자체에 의문을 던진다. 분류를 위해, 주석자가 동의하는 빈도를 계산할 수 있습니다. 등급 척도는 등급 사이의 평균 거리를 계산할 수 있습니다. 그런 다음이 원시 계약 통계를 계약 확률 (데이터를 무시한 두 주석 자 사이에서 얻은 계약 수준)과 비교해야합니다.

코헨의 카파 (Copper 's Kappa)는 이산 라벨링 작업에 대한 합의를 정량화하는데 널리 사용된다 (Cohen, 1960; Carletta, 1996),

[IR] Cohen's Kappa Coefficient(카파 상관계수)

분자는 관측 된 합의와 기회 합의 사이의 차이이며, 분모는 완전한 합의와 우연의 일치 사이의 차이다. 따라서, 모든 경우에 주석자가 일치 할 때 κ = 1이며, 주석자가 우연히 만날 때만 κ = 0이된다. κ가 "보통", "좋음"또는 "실질적인"동의를 나타낼 때 다양한 휴리스틱 스케일이 제안되었습니다.

참고로 Lee와 Narayanan (2005)은 구술 대화에서 정서의 주석에 대해 κ ≈ 0.45 - 0.47을보고하는데, "보통 합의"라고 표현한다. Stolcke et al. (2000)은 대화 행동의 주석에 대해 κ = 0.8을 보고하는데, 이것은 대화에서 각 차례의 목적을위한 레이블이다.

두 명의 어노 테이터가 있을 때 예상 확률 합의는 k가 레이블에 대한 합계이고 Pr (Y = k)가 모든 주석에 대한 레이블 k의 경험적 확률인 것으로 계산됩니다. 이 수식은 어노테이션이 무작위로 섞여있는 경우 기대되는 계약 수에서 파생됩니다. 따라서 이진 라벨링 작업에서 하나의 라벨이 인스턴스의 90 %에 적용되면 우연한 일치는 .92 + .12 = .82입니다.

4.3 Design decisions for text classification

4.5.2.2 Crowdsourcing

Crowdsourcing은 종종 분류 문제에 대한 주석을 빠르게 얻는 데 사용됩니다. 예를 들어, Amazon Mechanical Turk는 데이터 레이블 지정과 같은 "인간의 지능 업무 (Hit)"를 정의 할 수 있습니다. 연구원은 각 주석 세트에 대한 가격과 주석 달기에 대한 최소 자격 목록 (모국어 및 이전 작업에 대한 만족도)을 설정합니다. 상대적으로 훈련받지 않은 "군중 노동자"의 사용은 전문 언어 학자 (Marcus et al., 1993)에 의존했던 이전의 주석 작업과 대조됩니다.

그러나 crowdsourcing은 많은 언어 관련 작업에 대해 신뢰할 수있는 주석을 생성하는 것으로 밝혀졌습니다.

Crowdsourcing is part of the broader field of **human computation**