- 현재 경제 상황에 대한 논의 요약
  경제 상황, FX 및 국제 금융에 대한 MPB 회원들의 토론,
  금융시장과 통화정책.
- 통화 정책 결정에 관한 토론은 다음과 같은 견해를 기록합니다.
  개인 회원. • 통화정책
  심의 결과.

2005년 5월부터 2017년 12월까지의 MPB 회의록 파일(151분)을 한국은행 웹사이트에서 다운로드합니다.7 우리는 두 번째와 세 번째 섹션만 사용합니다. 패널(a)
(b) 그림 2에서 각 섹션에 대한 문장 수를 MPB 분 단위로 표시합니다.
시간이 지남에 따라. 글로벌 금융 위기 이후 회의록 길이가 늘어 났습니다.

뉴스 기사. "금리 (금리)" 라는 단어가 포함된 뉴스 기사를 모았 습니다.
2005년 1월부터 2017년 12월까지 Naver와 Infomax.8 이 뉴스
기사에는 경제 전반, 통화정책, 금융 시장, 한국은행의 향후 통화정책 기조에 대한 국민인식 등의 정보가 포함되어 있습니다. 우리는 만 사용
3대 통신사의 기사(기사 수 기준)는 원저자의 중복 기사가 많기 때문입니다. 최종 사용을 위한 뉴스 기사 수는 206,223개입니다. 그 중 42%(86,538) 가

연합인포맥스, 이데일리 33%(68,728), 연합뉴스 25%(50,957).
기사에서 머리글과 바닥글을 제거합니다. 그림 2의 패널 (c) 및 (d) 는
시간 경과에 따른 뉴스 기사의 수.

채권 분석가의 보고서. 우리는 또한 다음 두 가지 이유로 채권 애널리스트의 보고서를 사용합니다. 첫째, 채권 애널리스트 보고서는 통화 정책 및 채권 시장에 대한 전문가의 견해를 보여줍니다. 둘째, 우리는 비공식적인 글쓰기 스타일을
어휘. 일반적으로 채권 분석가는
언론인. 우리는 한국의 금융 정보 서비스 제공 업체 인 WIEfn에서 보고서를 입수했습니다.9 그림 2의 패널 (e)는 2005년 1월부터
2017년 12월.

우리 코퍼스는 크기가 크고 다양한 주제를 다룹니다. 그림 3은 다양한
Latent Dirichlet Allocation 방법을 사용하여 추출한 말뭉치의 주제,
주제 모델링 방법. 표 3은 주제의 상대적 빈도를 보여줍니다.

---

[7] 샘플 기간 동안 151분 동안 152회의 회의가 진행되었습니다. 글로벌 금융위기 당시 긴급회의는 회의록이 없었다. 의사록은 다음 링크에서 다운로드할 수 있습니다. http://www.bok.or.kr/portal/singl/crncyPolicyDrcMtg/listYear.do?mtgSe=A&menuNo
=200755

[8] https://news.naver.com, http://news.einfomax.co.kr

[9] https://www.wisereport.co.kr

[그림 3] 말뭉치의 토픽 워드클라우드

[표 3] 주제별 가중치 평균

아니.	주제명 1 외화	총	회의록 뉴스		보고서
2	금융정책 3 채권발행시장 1 4 통	5.24	11.20	5.94	3.75
화정책 5	채권발행시장 2 6 금융위	2.69	2.24	3.15	1.99
기 7	스왑시장 8 인플레이션 9 신	3.35	0.73	1.29	6.79
용등급 10	부동산 11 글로벌 국	3.81	12.56	4.47	2.20
채 12	거시안정성 13 부동산정책	2.79	1.32	2.67	3.08
14	유로존 15 경제 성장 16 머니마	1.79	1.03	2.09	1.36
켓 17	글로벌 주식시장 18 글로벌	4.30	3.05	4.02	4.82
통화정책 19	금융상품 20 기업가	3.32	10.56	2.68	3.89
치평가 21	자본요건 22 국채선물	1.42	0.38	0.93	2.26
23	국내주식시장 24 소호머니마켓	1.20	0.15	1.71	0.46
25	채권시장가격 26 캐리트레이드 27 정	2.23	0.40	1.34	3.78
부규제 28	펀드시장 29 기업구조조정 30	1.26	2.74	1.16	1.32
소비/소득 31	유동성 공급 32 정치 33 원	3.01	2.44	3.99	1.46
자재 가격 34	채권 투자 전략 35 주택 부	3.94	1.27	3.33	5.07
채 36	기업 신용 등급	5.02	13.60	4.58	5.19
		0.88	0.79	1.09	0.55
		2.11	0.37	2.74	1.20
		5.52	4.87	4.36	7.40
		2.34	0.22	3.42	0.74
		1.64	0.27	2.13	0.95
		0.77	0.42	0.48	1.27
		4.69	0.46	4.44	5.34
		0.81	0.17	1.21	0.23
		1.34	0.65	2.06	0.22
		1.01	0.33	0.52	1.82
		1.41	0.66	1.87	0.71
		1.43	0.87	1.75	0.95
		4.53	2.19	5.40	3.30
		1.43	1.66	1.59	1.16
		1.93	5.68	1.72	2.04
		0.61	0.21	0.46	0.88
		2.13	0.65	1.51	3.21
		1.61	0.90	1.54	1.78
		1.80	0.13	1.44	2.46
		5.99	8.03	7.81	2.95
		3.61	0.46	1.86	6.60

1.1. 텍스트 전처리

1.1.1. 사전 처리의 일반적인 단계

사전 처리 텍스트에는 토큰화 및 정규화가 포함됩니다. 토큰화는 긴 텍스트 문자열을 일반적으로 단어인 작은 조각 또는 토큰으로 분할하는 단계입니다. 토큰화는 품사(POS)를 통합할 수 있습니다.
태그 지정,

명사, 동사, 형용사 등과 같은 단어의 일부. 정규화는 텍스트를 단일 정식 형식으로 변환하는 프로세스입니다.
정규화에는 구두점 제거, 불용어 제거, 숫자를 상응하는 단어로 변환, 형태소 분석,

표제어화, 대소문자 접기.10
영어에 대한 일반적인 텍스트 전처리 절차는 (i) 모든 단어를 소문자로 변환, (ii) Porter(1980) 형태소 분석 알고리즘을 사용하여 숫자와 구두점을 제거하여 어근으로 굴절된 단어를 줄임(예: "increasing")을 포함합니다. "증가", "실업"에서 "실업"으로) 또는 표제어화(예: "더 나은"에서 "좋은"로) 및 (iii) 불용어 제거(예:

a, the, an, of, to 등).

1.1.2. 경제

한국어 텍스트를 숫자 표현으로 변환하는 데에는 몇 가지 문제가 있습니다(eg bag of words 및 word embedding). 첫 번째 문제는 간격과 관련이 있습니다. 영어와 달리 후치사는 공백으로 구분되지 않으며 간격 규칙도 엄격하게 준수되지 않습니다. 둘째, 수많은 외래어가 존재합니다.

외국인을 따르지 않는다
언어 표기 기준; 이러한 단어의 대부분은 분야에 따라 다릅니다. 셋째, 동의어에 대한 다양한 표기가 존재합니다(예: "인플레이션 "에 대한 인플레이션, "인
플레” , “물 가”). 이 문제는 n-gram을 사용할 때 중요할 수 있습니다.
다양한 표기법
동의어는 단어 조합의 수를 늘리고 빈도를 희석시킵니다.
n-그램. 넷째, 수많은 동사와 형용사가 불규칙적으로 활용됩니다.
불규칙한 컨쥬게이션은 또한 n-그램 모델에서 차원의 폭발을 악화시켜 극성 분류를 방해합니다. 간격의 첫 번째 문제는
현재 사용 가능한 한국어 형태소 분석기에서 비교적 잘 처리됩니다. 예를 들어 KoNLPy 를 사용할 수 있습니다.11 그러나 다른 문제는 그렇지 않습니다. 이를 위해 공동 저자인 Lee (2018) 가 개발한 eKoNLPy를 사용 합니다. eKoNLPy는 경제 및 특정 사전을 구성합니다.

금융 및 자체 형태소 분석기를 사용합니다.
두 번째 이슈와 관련하여 eKoNLPy는 사전 제공되는
인터넷에서 쉽게 구할 수 있는 경제용어사전에서 습득한 4,202개의 분야별 전문용어를 통해 경제와 금융을 완벽하게 지원
도메인 특정 용어(즉, 전문 용어 및 외래어).12 eKoNLPy는 POS 태깅을 위해 사용자 지정 용어 및 외래어를 사전에 쉽게 추가하는 기능이 있습니다. 3호는 eKoNLPy 사전에 동의어 1,325쌍을 미리 정의하고 동의어 대체 기능을 지원합니다.

동의어의 다양한 표기법을 처리합니다. 마지막 문제 즉 ,
형용사와 동사의 활용 형태소 분석 또는

---

[10] 불용어 제거는 "it", "the" 및 "etc"와 같은 불용어를 삭제하는 것입니다.
어간 추출은 단순히 어간을 세는 것입니다(예: "banking" 및 "banks"에 "bank" 사용).
표제어 추출은 단어의 변화된 형태를 그룹화하여 단일 항목으로 분석할 수 있도록 하는 것입니다. POS 태깅은 종종 원형 복원에 도움이 됩니다. 예를 들어, "톱"은 다음과 같을 수 있습니다.
동사 "보다"또는 명사의 과거 시제.

[11] KoNLPy 는 한국어 NLP용 Python 패키지입니다.
(http://konlpy.org/en/v0.5.1/ 참조).

[12] 네이버, 매경, 한경,

등.

원형화 우리말 단어의 불규칙 활용형을 정규화 하는 것은 형태소 분석보다는 원형 복원으로 해결할 수 있는데, 표제어 추출 은 단어의 형태학적 분석을 고려하기 때문입니다. eKoNLPy는 이러한 문제를 해결하기 위해 경제 및 금융 분야에서 자주 사용되는 1,291개의 형용사와 동사의 표제어 변환을 지원합니다.

eKoNLPy는 처음부터 경제 분석을 위한 텍스트 마이닝을 목표로 개발되었기 때문에 경제 분석에서 KoNLPy에 비해 우수한 성능을 기대합니다. 예를 들어 다음을 고려합니다.

문장:

> "한국은행이 12일 금융통화위원회(금통위) 회의를 열고 기준금리를 현행 연 1.50로 동결했다."

우리는 eKoNLPy가 "금융통화위원회 (Monetary Policy Board)"와 "금통위 (MPB) "를 성공적으로 인식하고 있음을 발견했습니다. 범용 사전 을 사용하는 KoNLPy는 불가능합니다.13 채권 시장에 대해 다음과 같은 문구를 고려합니다.

> "금리 박스권 상단 상향과 일드 커브 완만한 스티프닝 전망."

eKoNLPy는 '일드 커브(yield curve)'와 '스티프닝 (steepening)'을 인식하지만 KoNLPy는 인식하지 못한다.

1.2. 기능 선택

모든 단어가 의견을 표현하는 데 사용되는 것은 아니므로 대상 목록에 단어 또는 구를 제한하는 기능 선택 수행 의견을 표현하는 말. 단어를 제한하면 용어(단어) 벡터의 차원이 줄어들어 처리 속도가 빨라집니다. 게다가, 단일 단어는 종종 문맥을 잃습니다. 예를 들어, "회복"이라는 단어는 단독으로 긍정적인 메시지를 전달하는 것처럼 보이지만 "부진한 회복"은 그렇지 않습니다.14 긍정적일 때와 부정적 일 때

---

[13] KoNLPy의 결과는 다음과 같습니다.

"한국은행/NNP," "이/JKS'" "12/SN," "일/NNBC," "금융/NNG," "통화/NNG," "위원회/NNG,"
"(/SSO," "금/NNG," "통/NNG," "위/NNG," ")/SSC," "회의/NNG," "를/JKO," "열/VV,"
"고/EC"
"기준/NNG," "금리/NNG," "를/JKO," "현행/NNG," "연/NNG," "1/SN," "./SY," "50/SN,"
"%/SY," "로/JKB," "동결/NNG," "했/XSV," "다/EF," "./SF"
The result from eKoNLPy is as follows:
"한국은행/NNP," "이/JKS," "12/SN," "일/NNG," "금융통화위원회/NNG," "금통위/NNG,"
"회의
/NNG," "를/JKO," "열/VV," "고/EC," "기준금리/NNG," "를/JKO," "현행/NNG,"
"연/NNG,"
"1/SN," "./SY," "50/SN," "%/SY," "로/JKB," "동결/NNG," "했/XSV," "다/EC."

약어 는 NNP, JKS 및 SN과 같은 POS 태깅을 위한 것입니다 . 예를 들어 NNG, JKS 및 SN은 일반을 의미합니다.
명사, 주격격 후치사, 수. 의 표 참조
자세한 내용 은 부록 B를 참조하십시오.

[14] Apel과 Grimaldi(2014)는 명사와 "더 높은 인플레이션" 또는 "더 느린 성장"과 같은 형용사 분류.

실업률을 낮춰라'는 바이그램(bi-gram) 문구 등의 단어가 결합돼 있어 정서를 가늠하기 쉽지 않다. 따라서 우리는 이 문제를 해결하기 위해 n-gram을 사용합니다.15 그러나 n-gram의 길이를 늘리는 것은 장단점이 있습니다. 매우 긴 n-그램(예: 10그램)을 사용하면 샘플에 과적합되는 문제에 빠질 수 있습니다. 어휘집은 대상 문서에 매우 구체적이기 때문에 이러한 어휘집을 뉴스 기사나 전문가의 글과 같은 다른 유형의 문서에 적용하는 것은 어렵습니다. 게다가 n-gram에서는 차원의 저주가 발생합니다 .16

동일한 기능 을 가진 n-gram을 볼 확률 이 작아집니다. 이러한 차원의 폭발은 또한 메모리 크기 및 처리 속도와 관련된 계산 문제를 야기합니다.

이 트레이드 오프를 해결하기 위해 추가 규칙과 함께 n-gram의 n을 5로 설정 했습니다 .17 차원의 폭발 을 피하기 위해 단어의 품사 태그를 제한하여 n-gram을 형성할 때 제한된 단어 세트를 사용합니다. 명사에
(NNG), 형용사(VA, VAX), 부사(MAG), 동사(VA), 부정.18
분류의 정확도를 높이고 다중 계산을 피하기 위해 각 문장에서 여러 개의 중첩 n-gram이 발견될 때 가장 높은 n-gram만 고려합니다. 발생하는 n- 그램 도 드롭합니다.

15회 미만.19
최종 단어 집합은 2,712개의 단어로 구성되었으며 결과를 얻습니다.
73,428n-그램. 특히, 우리의 n-gram은 1에서 5-gram까지 다루기 때문에 자연스럽게 단일 단어(1-gram)를 포함합니다.20 다음 단계는 문장의 감정을 측정하기 위해 이러한 n-gram의 극성을 분류하는 것입니다.
서류.

---

15 Picault와 Renault(2017)는 샘플에서 최소 두 번 나타나는 n-gram(1-gram에서 10-gram까지)을 고려하여 분야별 어휘를 정의합니다. 그들은 ECB 입문서 에서 발음되는 모든 문장을 수동으로 분류한 후 어떤 범주의 문장(dovish, neutral, hawkish 또는 positive, negative, neutral)에 속할 확률을 계산하여 n-gram의 극성을 분류합니다 . n-gram을 각 클래스에서 확률이 0.5 이상인 것으로 제한함으로써 최종 필드별

lexicon은 34,052개의 n-gram으로 구성되어 있습니다.

16 텍스트는 매우 고차원이지만 희박한 벡터로 표현되기 때문에 문서 전체에서 중요한 변형을 유지하면서 차원을 줄이는 것은 어려운 일입니다.
텍스트에서 n 단어 의 연속 시퀀스인 n-gram의 도입으로 이 문제는 더욱 악화되었습니다. 1,000개의 고유 단어 말뭉치로 바이그램 모델에 필요한
1,0002개의 값; 트라이그램 모델에는 1,0003이 필요합니다 . 등등.

17 Hutto와 Gilbert(2014)를 포함하여 수많은 연구에서 n-gram 접근 방식이 감정 분석의 성능을 향상시킨다고 보고합니다. 일반적으로 그들은 bigram을 사용하여
5그램. Dey, Jenamani 및 Thakkar(2018)는 영역 독립적인 n-gram 정서 를 생성하기 위해 최초의 완전 자동 점수 계산 알고리즘을 제안합니다.
사전, 그들은 계산 부담 때문에 최대 트라이그램을 사용합니다.

18 부록 B는 POS 태깅을 위한 eKoNLPy 태그 세트를 보여줍니다.

19 감정 분석을 위한 효과적인 의견 전달 기능인 n-그램의 관점에서 이상적인 n-그램에는 다음 요소가 포함되어야 합니다.

(즉, 감정이 느껴지는 실체), 어떻게 (즉, 영향의 방향 또는 정도)
정서적 상태) 및 부정이 있는 경우.

20 또한 n-gram에서 n 의 민감도를 확인 하고 결과가 민감하지 않음을 확인합니다.
n (= 2, 3, 4 및 5) 의 선택에 . 우리는 또한 n 이 높으면 정확도 측면에서 샘플 내 성능이 증가하고 샘플 외 성능이 낮아진다는 것을 발견했습니다.
극성 분류. 이 결과는 높은 n 이 n-gram을 높게 렌더링 함을 시사합니다.

문서별.

1.3. 극성 분류

Harvard-IV와 같이 잘 알려진 극성 단어 목록이 없는 경우
또는 LM 사전에서 선택한 기능의 극성을 분류해야 합니다.
(우리의 경우 n-grams) 자체적으로.21 여러 범주의 극성 분류가 존재할 수 있습니다. 첫 번째는 사람의 개입 이 필요한지 여부에 따라 감독 대 감독되지 않은(자동화된) 접근 방식 을 나타냅니다. Google Cloud Sentiment Analysis API는 분류자가 방대한 양의 문서에 대해 훈련되는 지도 분류의 예입니다. 감독되지 않은 접근 방식의 예는 단어와 극성 프로토타입 간의 유사성을 측정하기 위해 점별 상호 정보(PMI)를 사용하는 의미론적 방향입니다.22 두 번째는 기계 학습 대 어휘 기반 방법을 나타냅니다. 전자는 극성 정보로 주석이 달린 훈련 코퍼스를 사용하고 후자는 극성 어휘집을 사용합니다. 어휘 기반 접근법에서 극성 어휘를 얻기 위한 세 가지 방법 , 즉 수동, 사전 기반 및

말뭉치 기반. 수동 방법 은 시간이 많이 걸리고 인적 오류가 발생하기 쉽습니다.23 사전 기반 접근 방식은 시드 단어에서 시작하여 사전을 검색하여 동의어와 반의어를 결정합니다.

이 접근 방식에는 WordNet과 같이 잘 구성된 어휘 데이터베이스가 필요합니다.
또는 thesaurus.24 이 접근법의 단점은

---

[21] 유니그램(단일 단어)의 경우 가장 오래된 것은 General Inquirer(Stone, Dunphy 및 Smith, 1966) Harvard IV-4로도 알려져 있습니다. 후자는 긍정적인 전망의 1,915단어와 부정적인 전망의 2,291단어를 포함하여 다양한 범주의 단어 목록을 가지고 있습니다. ~ 안에
재무적 맥락에서 부정적인 단어는 감정 분석에 사용됩니다(Tetlock, 2007). ㅏ
금융 문헌에서 널리 사용되는 단어 목록은 Loughran and McDonald(2011)의 단어 목록입니다.
범주별로 단일 단어 목록이 있습니다(부정적, 긍정적, 불확실성, 소송, 모달 및 제약). 그들의 연구는 LM 사전이 좋은 것을 가지고 있음을 나타냅니다.
재무 지표와의 상관 관계. LM 사전 -아ㅋ 사용 가능
https://sraf.nd.edu/textual-analysis/resources/ .

[22] 의미론적 방향은 컴퓨터 언어학의 개념이며 다음을 정의합니다.
의 위치
두 개의 반대 개념 사이의 단어 또는 단어 문자열. PMI는 확률 이론을 기반으로 두 랜덤 변수 간의 유사성을 정량화하는 기법입니다.
PMI를 사용하여 두 어휘 사이 의 유사성은 w1 과 w2 가 다음 과 같이 측정됩니다 .
고려중인 두 개의 어휘. Lucca 및 Trebbi(2011) 및 Tobback, Nardelli 및 Martens(2016)는 이 의미론적 방향 PMI(SO-PMI) 방법을 사용하여 FOMC 성명 및 ECB 언론 성명의 감정을 측정합니다. 그들은 Google 검색 을 사용하여 열반적으로 통화 정책의 정서(비둘기적 또는 매파적)와 관련된 단어와 함께 문서에서 단어 또는 문자열의 동시 발생을 계산하여 정서 지표를 측정합니다.

[23] 따라서 작은 집중 단어 목록만으로 특정 작업에 사용됩니다.
Apel과 Grimaldi(2014), Bennani와 Neuenkirch(2016) 의 사례처럼 . [24]
WordNet®은 대규모 영어 어휘 데이터베이스입니다. 명사, 동사, 형용사 및 부사는 인지적 동의어(synsets) 집합으로 그룹화되며, 각각은
뚜렷한 개념.

$$PMI(w, w) = \log \frac{P(w1, w2)}{P(w1)P(w2)},$$

필드별 극성 단어를 결정할 수 없음. 말뭉치 기반 방법
큰 코퍼스 에서 시드 단어와 함께 함께 발생하는 패턴을 검색하여 극성 단어를 찾습니다 . 이 접근 방식은 해당 도메인 의 코퍼스를 사용하여 필드 및 컨텍스트별 감정 단어와 그 극성을 찾을 수 있다는 점에서 큰 이점이 있습니다. 따라서 말뭉치 기반 접근법은

경제 또는 금융 분야에 가장 적합합니다.
다른 의미가 널리 퍼져 있습니다.
우리는 n-gram의 극성을 두 가지 방식으로 분류합니다. 하나는 머신러닝을 이용해 시장 정보로부터 극성을 분류하는 시장접근법이다.
다른 하나는 단어를 사용하여 극성을 분류하는 말뭉치 기반 접근 방식입니다.
(우리의 경우 n-gram) 임베딩 및 시드 단어, 우리는 이를 어휘 접근 방식이라고 합니다.25

1.1.3. 시장 접근

Söderlind and Svensson (1997) 의 조사에서 제시된 바와 같이 자산 가격에서 시장 기대치에 대한 정보를 추출하려는 수많은 시도가 있었습니다 . 금융시장이 효율적이고 자산가격이 금융시장의 정보를 반영하는 정도까지 정보를 추출할 수 있다.

그렇다면 텍스트에서 정보를 추출할 수 없는 이유는 무엇입니까? 이러한 고려 사항을 염두에 두고 시장을 사용하여 기능의 극성을 분류합니다.
정보를 제공하고 이를 "시장 접근 방식"이라고 합니다.

단어의 상대적 가중치를 결정하기 위해 Jegadeesh와 Wu(2013)는 다음을 사용합니다. 단어를 설명변수로, 주식수익률을 종속변수로 사용합니다. 단어의 계수가 양수이고 크면 단어의 가중치가 높습니다. 즉, 시장의 반응을 추정하여 상대적 가중치로 사용하였다. Jegadeesh and Wu(2013) 의 한 가지 장점은 주관적 판단에 의존하지 않는다는 점이다. 전자는 Harvard IV-4 또는 LM 사전과 같은 알려진 단어 목록에서 시작하고 회귀 기반 접근 방식을 사용하여 용어의 상대적 가중치(부호 및 크기)를 결정하기 때문에 그들의 작업 은 우리의 시장 접근 방식과 구별됩니다 . 또한 n-gram이 아닌 uni-gram을 사용합니다. 우리의 시장 접근법은 또한 시장 정보를 사용하여 기능의 극성을 결정합니다. 그러나 n-gram을 추출하려면

큰 코퍼스에서 극성을 분류하고 기계 학습을 사용합니다.
방법.

시장 접근 방식을 위해 간단한 확률 분류기인 NBC(Naive Bayes Classifier)를 사용합니다. NBC는 매우 간단하지만 서포트 벡터 머신을 포함한 고급 방법 으로 여전히 경쟁력이 있습니다 .26 NBC는 주어진 모든 기능이 독립적이라고 가정하기 때문에 순진하다고 합니다.

계급(매파적이거나

//wordnet.princeton.edu.

[25] 단어 임베딩은 NLP의 언어 모델링 및 기능 학습 기술 집합에 대한 총칭으로, 어휘의 단어나 구가 실수 벡터에 매핑됩니다. 개념적으로는 수학적

단어당 차원이 하나인 공간에서 차원이 낮은 연속 벡터 공간으로 임베딩(https://en.wikipedia.org/wiki/Word_embedding).

[26] NBC 및 서포트 벡터 머신에 대한 자세한 내용은 다음 링크를 참조하십시오: https://www.ocf.berkeley.edu/~janastas/supervised-learning-with-text-II-03-16-Lecture.html#1.

우리의 경우 dovish). NBC는 기능 선택 도구가 아니지만 이러한 독립성 가정으로 인해 각 기능의 조건부 확률을 극성 점수로 사용합니다.

NBC를 포함한 기계 학습 방법은 대부분 지도 방식이며 레이블이 지정된 훈련 문서의 존재에 의존합니다. 교육 문서는 일반적으로 공개된 검토를 통해 얻을 수 있습니다. 그렇지 않으면 여러 전문가가 교육 문서에 수동으로 레이블을 지정해야 합니다. 전자는 통화 정책에 사용할 수 없습니다. 후자는 노동력과 비용 집약적이며 전문가의 판단에 따릅니다. 이러한 문제를 회피하고 금융 시장의 정보를 활용하기 위해 우리는 코퍼스의 뉴스 기사와 보고서 가 발표된 날에 한 달 동안의 콜 금리 변화가 긍정적(부정적) 이면 매파적(dovish)으로 분류합니다.27

우리는 레이블이 지정된 문장(400만 개 이상의 문장)을 9:1 비율로 훈련 세트와 테스트 세트로 무작위로 나눕니다.28 각 문장의 기능 으로 5-그램(1~5그램)을 사용 하여 분류기를 사용하여 정확도를 확인합니다. 훈련된 NBC는 주어진 클래스(매파적/ 비두적적)에 대한 각 기능의 조건부 확률을 산출하며, 이를 기능의 극성 점수로 사용합니다.

$$\text{수} = p(\text{특징}|\text{비둘거}) \frac{p(\text{특징}|\text{매파})}{\text{극성 점}} = \frac{p(\text{피처 \& 매파}) / p(\text{매파})}{(1)\ p(\text{피쳐 \& 비둘기}) / p(\text{비둘기})}.$$

n-gram 이 "dovish" 문서에 비해 "hawkish " 문서 에서 더 자주 나타나는 경우 대략적으로 "hawkish"로 레이블이 지정됩니다 .29

무작위 샘플링과 확률적 분류기를 사용한다는 점을 감안할 때 모든 교육은 각 클래스의 사후 확률이 다릅니다. 좋은 예측 성능을 얻기 위해 이 절차를 30회 반복하고 극성 점수의 평균을 최종 값으로 사용하는데, 이를 기계 학습에서 배깅(bagging)이라고 합니다.30 어휘집의 직접적인 성능 측정은 아니지만 NBC의 평균 정확도는 다음과 같습니다. 86% (긍정적 정확성: 90%, 긍정적 재현율: 84%, 부정적 정확성: 82%, 부정적 재현율: 88%).

---

27 그만큼 실제 무의미한 움직임을 제외하려면 ±3dp 입니다.
임계값은

28 훈련, 테스트 및 검증의 세 가지 세트를 사용하도록 제안할 수 있습니다.
섹션 3.4.3에서 샘플 외부 문서로 극성 분류를 평가한다는 점을 감안할 때 유효성 검사 세트가 필요하지 않으며 샘플 외부 평가는 매우 엄격합니다.

29 또는 다음을 고려할 수 있습니다.

$$\text{수} = p(\text{도비시}|\text{특징}) \frac{p(\text{매파}|\text{특징})}{\text{극성 점}} = \frac{p(\text{특징 \& 매파})}{\text{\& 비둘기파)}}\ p(\text{특징}).$$

우리의 경우 결과 때문 p(매파적)= 0.53 및 p(비두적적) = 0.47 , 이 두 공식은
입니다.

비슷한

[30] 배깅은 앙상블의 각 모델에 동일한 가중치를 부여하는 부트스트랩 집계의 약어입니다.