텍스트 마이닝을 활용한 금융통화위원회 의사록 분석 요약본

1. text mining을 이용하여 2005년 5월부터 2017년 12월 사이의 232,658개의 통화 정책 결정에 대한 정량적인 정보를 추출

* 한국은행 금융통화위원회 회의록의 양적지표를 질적지표로 변환함으로서, 회의록의 느낌과 톤을 측정하고, MPB communication이 이용 가능한 거시경제 데이터에 포함되지 않은 추가 정보를 전달하는지 여부를 조사

1. ‘저실업’이나 ‘경기회복 부진’과 같은 긍정적이고 부정적인 단어를 조합한 단어의 어조나 정서(unigram)을 결정하기 쉽지 않기 때문에 n-gram사용

* 미리 결정된 seed words에 대한 근접성에 근거하여 극성을 결정

1. 자체의 자연어 처리 도구인 eKoNLPY를 사용하여 field-specific의 비한국어 외래어, 동사와 형용사의 불규칙한 결합 등 한국어와 관련된 어려움을 해소

* eKoNLPy는 GitHub에서 공개

1. word2vec의 방법을 비교하고 word2vec의 성능을 보여줌으로써 향후 연구방향에 대한 유용한 지침을 제공
2. word2vec은 반어법을 종종 유사한 단어로 분류하는 문제를 안고 있어, 이 문제를 N-gram Embedding으로 해결

첫째, 문맥을 고려하여 텍스트의 미묘한 점을 포착하기 위해 N-gram을 사용한다.

둘째, 우리는 경제학과 금융분야에서 Text Mining을 위해 특별히 고안도니 현장 고유 사전을 사용한다. 셋째, 어휘적 접근법 외에 시장접근법을 바탕으로 심리지수를 구성한다.

**\*감성 분석**

분석하고자 하는 것에 대한 말뭉치 준비,

1) MPB 회의록

-2005년 5월부터 2017년 12월까지의 기간 동안 231,699건의 문서를 수집하는데, 여기에는 151분간의 MPB 회의, 206,223개의 뉴스기사, 26,284개의 채권 분석 보고서가 포함

-2005년 5월부터 2017년 12월까지의 MPB분 파일을 한국은행 웹사이트에서 다운로드, 두 번째와 세 번째 섹션만을 사용

2)뉴스 기사들

네이버와 인포맥스로부터 2005년 1월부터 2017년 12월까지 ‘금리’라는 단어를 포함한 뉴스기사를 수집

최종적으로 사용할 뉴스기사의 수는 206,223개 이다. 이 중 42%(86,538)은 연합 Infomax, 33%(68,728)은 EDAILY, 25%(50,957)로 집계, 기사에서 헤더와 바닥글을 제거

3)채권 분석가의 보고서

채권 분석 보고서가 통화정책과 채권시장에 대한 전문가들의 견해를 나타낸다는 것

한국의 금융정보서비스 제공업체인 WIEfn으로부터 보고서를 받음. 2005년 1월부터 2017년 12월까지의 보고서

주제 모델링 방법인 LDA(Latent Dirichlet Allocation) 방법을 사용하여 추출

\*텍스트 전처리,

한국어 자연어 처리 Library (eKoNLPy)

eKoNLPy는 4,202개의 field-specific 용어들이 준비

eKoNLPy는 사전에 1,325 쌍의 동의어를 정의하고 동의어 대체 기능을 지원

**\*Feature Selection (특징선택)**

n-gram의 n을 5로 설정

차원의 폭발을 피하기 위해 단어의 부분어 태그를 명사(NNG), 형용사(VA, VAX), 부사(MAG), 동사(VA), 부정으로 제한하여 n-그램 형성에 설정된 한정어를 사용한

15번 이하로 자주 발생하는 n-그램도 떨어뜨린다

최종 단어 세트는 2,712개의 단어로 구성되어 있으며, 우리는 그 결과 73,428개의 n-grams를 얻는다.

1그램에서 5그램까지 다루기 때문에, 우리의 n그램은 자연스럽게 단어들(1그램)을 포함

Polarity Classification

인간의 개입이 필요한지에 따라 감독된 접근 방식과 감독되지 않은 접근 방식(자동화)

분류에 의견어를 사용하는지에 따라 기계학습 기반 대 어휘 기반 방법

n-그램의 극성을 **두 가지 방법**으로 분류

**(1)기계학습으로 시장정보로부터 극성을 분류하는 시장접근법**

시장접근법에 대해서는 단순한 확률론적 분류자인 Na ıve Bayes 분류기(NBC)를 사 용

무작위로 라벨이 붙은 문장(4백만 문장 이상)을 훈련 세트와 9:1 비율로 설정된 시 험으로 나눈다

문장별로 5그램(1그램부터 5그램까지)을 형상으로 삼아 분류기를 훈련시키고 그 정확도를 확인한다

훈련된 NBC는 우리가 특징의 극성 점수로 사용하는 클래스(hawkish/dovish)를 주어진 각 특 징의 조건부 확률을 산출한다

무작위 표본 추출과 확률론적 분류자를 사용하기 때문에, 모든 훈련은 각 등급의 다 른 후방 확률을 산출한다. 더 나은 예측 성능을 얻기 위해, 우리는 이 절차를 30번 반복하고 극성 점수의 평균을 최종 점수로 사용한다. 그것은 기계학습에서 bagging이라고 한다

NBC의 평균 정확도는 86%(긍정 정밀 도:90%, 긍정적 리콜:84%, 부정적 정밀도:82%, 부정적 리콜:88%)이다. 극성 점수가 1보다 크면(낮음) lexicon의 극성을 매파적(dovish)으로 분류하고, 강도는 1.3 을 임계값으로 사용하는 회색 영역의 lexicon을 제외한다.3) 최종 lexicon 수는 매족 1만8685개, 비둘기 2만1280개

Lexical Approach

어휘적 접근은 직관적인 관 찰에 기초한다: 만약 두 단어가 같은 맥락에서 자주 함께 나타나면, 그들은 같은 극성을 가 질 가능성

알 수 없는 단어의 극성은 다른 단어와 함께 발생하는 상대적 빈도를 계산하여 결정 할 수 있다. 이것은 PMI(Pointwise Mutual Information)의 개념을 사용함으로써 이루어질 수 있다.

SO-PMI(PMI에서 Semantic Orientation)

반어법(antonym)을 인식하지 못하는 경우가 있는데, 이는 동시 발생에 근거하여 극성을 판단하기 때문

->단어 내장 대신 Zhao, Liu, Li, Li, and Du (2017)의 ngram2vec를 사용

결과는 종자의 선택에 의해 영향

->최첨단 기술인 해밀턴 외(2016년)에 의해 SentProp 프레임워크를 채택

**(2)어휘적 접근법이라고 부르는 단어(우리 경우 n-gram)와 종자 단어를 사용하여 극성을 분류하는 말뭉치 기반 접근법**

말뭉치의 전체 232,658개의 문서를 이용하여 ngram2vec를 훈련

훈련에 사용하는 매개변수는 중심어용 5그램, 문맥어용 5그램, 창 크기 5개, 음의 표본 추출 크기 5개, 벡터 표현용 300차원

말뭉치는 최소 주파수 한도가 25인 344,293개의 고유 n-그램이 있어 410,902,512쌍 의 n-그램(크기 21.7GB)을 산출

시장접근법과 마찬가지로 극성점수가 1보다 클 경우(낮을 경우) 렉시콘의 극성을 매파적(도 비적)으로 분류하고 1.1의 강도를 임계값으로 사용하는 회색 영역의 렉시콘을 제외한다. 최종 렉시콘 수는 매파 11,710개, 비둘기 12,246개다.

시장과 어휘적 접근법이 유사한 결과를 제공하는지 보기 위해, 우리는 공통의 n-그램의 수 를 세어본다. 시장접근법에서 39,965개의 n-그램과 어휘적 접근에서 23,956개의 n-그램으로 볼 때, 14,154개의 공통 n-그램이 있다. 이 중 극성이 같은 n그램은 9791개(일반 n그램의 69%)이다.

**\*Evaluation**

어휘 분류의 정확성을 확인

2009년 5월부터 2018년 1월까지의 문서로 우리는 2,341개의 문장을 매파, 중립, 도회지로 수동으로 라벨을 붙인다

분류의 일관성을 확인하기 위해 무작위로 선택된 Na Bayve Bayes 분류기를 교육한다. 매파적이고 독설적인 문장의 60%와 나머지 문장으로 시험한

30회 반복하면 분류자의 평균 정확도는 약 86%로, 우리가 생각하기에 평준화 정확도

시장접근법에 의해 발생되는 어휘소의 경우 정확도는 68%(positive precision: 63%, positive recall: 75%, negative precision: 74%, negative recall: 62%)이다. 어휘적 접근에 의해 생성되는 어휘소의 경우 정확도는 67%(positive precision: 69%, positive recall: 71%, negative precision: 65%, negative recall: 62%)이다.

**\*Measuring Sentiments (**문장 및 문서의 정서를 측정하는 과정)

각 문장의 매파적 특징과 도회적 특징(n-gram)의 수를 기준으로 문장의 어조를 계산

렉시콘을 손에 쥔 채, 마지막 단계는 우리의 목표 문서의 톤을 측정

매파적 특징과 도회적 특징(n-gram)의 수를 기준으로 문장의 어조를 계산

Measures of MP Sentiment(MP감정 측정) 제3장에서 논의하는 방법론을 바탕으로 한은 MPB 회의록의 정서(또는 톤)를 포착하는 어휘 기반 지표를 개발한다: 전자는 시장접근법을, 후자는 어휘접근법을 사용한다. 그림 4는 의 시계열과 한국은행의 정책금리, 기타 경제불확실성의 측정을 보여주고 있다. 그림 4의 패널(a)은 의 시계열을 보여준다. 그들은 서로 가깝게 움직인다. 두 지표 사이의 상관 계수는 0.85

우리는 현장 **고유 한국어 사전과 n-그램으로 통화정책의 정서를 계량화하는 텍스트 기반의 지표를 개발**한다. 우리는 **우리의 지표가 현재와 미래의 통화정책 결정을 설명하고 다른 지표에 비해 더 나은 성과를 거두는 데 도움이 된다는 것을 보여준다**. 우리는 또한 현장 고유 사전과 원본 한국어 텍스트를 사용하는 것이 중요하다는 것을 보여주는데, 이것은 이 분야의 향후 연구에 도움이 될 것이다. 우리의 경험적 결과는 미래의 연구 장소를 암시한다. 첫째, **우리 지표들이 한은의 정책금리나 거시경제변수와 비교해서 어떤 종류의 정보를 갖고 있는지(또는 그렇지 않은지) 살펴보는 것이 중요**하다. 한은의 정책금리를 통화정책심리 지표의 한계변수나 잠재변수로 해석한다면 통화정책의 효과를 분석하는 표준 VAR시스템이나 DSGE모델에 우리의 대책을 주입하는 것이 흥미로울 것이다. 또 다른 방향은 한센과 맥마흔(2016년)이 될 것이다. 그들은 연준 성명서로부터 경제 상태와 선도적 지침에 관한 두 가지 별도의 지표를 작성하고, 이 두 가지 차원의 중앙은행 통신이 금융시장과 실제 변수에 어떻게 영향을 미치는지 검토하기 위해 요인 증강 VAR(FAVAR)을 사용한다. 둘째, **우리의 조치는 선도적 지침을 포함한 중앙은행 통신의 효과를 평가하는데 사용**될 수 있다. 한은의 의도와 상관없이 통화정책의 향후 결정을 설명하는 데 회의록에 근거한 우리 측 지표가 도움이 된다. Page34. 예를 들어 한은 발표 전후에 고주파 MP심리지표를 만들면 발표 전후에 '서바이벌'을 측정하는 데 활용할 수 있다. 게틀러와 카라디(2015년)는 연준의 통화정책 발표로 인한 놀라움을 측정하기 위해 발표 전 10분에서 발표 후 20분까지 연방기금 선물금리 변화를 이용한다. 국내에는 정책금리 선물 같은 것이 없다는 점을 감안하면 문자메시지표가 괜찮은 대안이 될 것이다. 셋째, **우리의 방법론은 거시경제 불확실성, 미래의 통화 정책 기조에 대한 대중의 기대, 주식 시장 정서 등을 측정하기 위해 다른 지표들을 건설하는 데 쉽게 적용**될 수 있다. 이러한 조치의 변화가 자산가격이나 실제 변수에 어떻게 영향을 미치는지 조사할 수 있다. 우리는 이 분야에서 해야 할 일이 더 많다는 것을 잘 알고 있지만, 우리의 연구가 출발점으로서, 텍스트 마이닝 접근법이 통화 정책을 분석하고 목적을 달성하는 한국은행과 연구자들의 도구 상자에 유용한 추가가 될 수 있다는 것을 보여주기를 바란다.