經濟學硏究 제71집 제2호 57-96
DOI : 10.22841/kjes.2023.71.2.002

인터넷 댓글의 정책여론 대표성 평가: 가상자산 과세의 사례*

김 수 현** · 배 진 수***

논문초록

본고는 가상자산 기타소득세 신설이라는 조세정책의 사례를 통해 인터넷 기사 댓글 내용이 조세정책에 대한 여론을 대표하는지 여부를 평가하였다. 조세정책과 관련된 인터넷 기사는 누구나 쉽게 접할 수 있고 그에 대한 의견도 댓글로 자유롭게 표현할 수 있다. 따라서 기사에 대한 댓글들이 정책에 대한 국민들의 여론을 나타낸다고 볼 수도 있을 것이다. 본 연구에서는 세 차례에 걸친 무작위 표본추출을 통한 주관식 설문조사를 바탕으로 트랜스포머 모형을 학습하여 가상자산 기타소득세 신설 정책에 대한 긍정/부정을 분류할 수 있는 감성분류기를 구축하고 해당 조세정책에 대한 기사 댓글에 대해 긍정/부정 비율을 구하였다. 그 결과 기사의 댓글에 나타난 정책반응은 부정의견의 비율이 전체 댓글의 80~90%로 나타났다. 이는 무작위 표본추출을 통한 주관식 설문조사의 부정비율 약 30%에 비해 상당한 수준의 편의를 보이는 것을 알 수 있다. 본 연구의 결과는 향후 조세정책뿐만 아니라 정부의 정책에 대한 국민들의 반응을 수렴하는 과정에서 기사의 댓글을 근거자료로 삼을 것인지 여부를 판단하는데 시사점을 준다.

핵심 주제어: 댓글, 인터넷 여론, 조세정책, 감성분류, 딥러닝
경제학문헌목록 주제분류: C8, H2, H3

투고 일자: 2023. 2. 15. 심사 및 수정 일자: 2023. 5. 4. 게재 확정 일자: 2023. 5. 23.

* 본 논고의 구성과 질을 크게 향상 시킬 수 있도록 세심한 논평을 해 주신 두 분의 익명 심사자 분들께 감사의 말씀을 드린다. 본 연구를 위해 귀중하고 유익한 조언을 주신 한국조세재정연구원의 강동익, 고창수, 김정환 부연구위원, 2023 경제학 공동학술대회 한국경제학회 경제학연구 특별세션 참가자들께 감사의 말씀을 전한다. 또한 모형 관련 기술적 조언을 주신 제주한라대학교 이영준 교수께도 깊은 감사의 말씀을 전한다. 본 연구는 2022년도 한국조세재정연구원의 기본연구과제 「텍스트 분석을 이용한 조세정책에 대한 인식 연구」의 일부 내용을 수정·보완한 것이며, 2023 경제학 공동학술대회의 경제학연구 특별세션(빅데이터, 비정형 자료, AI를 활용한 응용 경제 연구)에서 발표한 논문이다. 남아있는 오류는 저자의 책임임을 밝힌다.

** 제1저자, 전남대학교 경제학부 조교수, e-mail: soohyon.kim@jnu.ac.kr
*** 교신저자, 한국조세재정연구원 부연구위원, e-mail: jsbae@kipf.re.kr

Ⅰ. 서 론

“세상에 좋은 세금이란 없다”는 윈스턴 처칠의 말처럼 세금은 국가재정에 꼭 필요하지만 국민들에게는 반갑지만은 않은 부담일 것이다(박지웅 외, 2018). 그렇기에 조세부담을 증가시키는 세법의 개정이나 신규 세목 신설은 국민적인 반대와 저항에 부딪히기 쉽다. 이러한 저항이 언론의 관심과 여론을 통해 형성되는 경우 계획된 조세정책이 수정되거나 폐지되는 경우도 종종 발생[1] 한다.

조세가 정치와 밀접한 연관이 있다는 점에 대해서는 이견의 여지가 없을 것이다(Gould and Baker, 2002). 국민의 대표기관인 의회의 승인이 없이는 국민에게 과세할 수 없다는 조세법률주의[2]가 의미하는 바와 같이 조세정책은 국민의 선호를 잘 반영하는 경우에만 그 정당성을 확보할 수 있다. Torgler(2007)는 스위스와 미국을 대상으로 한 연구에서 국민발안(legislative initiatives)이나 국민투표(legislative referendum)처럼 직접민주주의 수단이 더 발달된 지방정부일수록 더 높은 납세의식(tax morale)을 발견할 수 있다고 주장했다. Daude(2012) 연구에서는 민주주의가 최선의 정치제도라고 생각하는 사람일수록 탈세를 정당화 하지 않는다는 결과를 얻었다. Jun et al. (2015)은 실험 연구를 통해서 세율이 다수결의 원칙으로 결정되는 경우 독재의 결과로 정해질 때보다 더 높은 납세순응도를 보인다는 것을 확인하였다. 이러한 연구들은 조세정책의 수립이 민주적 의사결정과정을 통해 이루어진다면 납세 행정의 측면에서도 효율성을 확보할 수 있다는 점을 시사한다.

국내의 조세정책은 정부가 발의하는 세법개정안을 중심으로 이루어졌으나 법안 실명제가 도입된 2000년도 이후 의원발의안의 수가 증가하기 시작했으며 2010년도 이후에는 이러한 현상이 가속화되고 있다(국회예산정책처, 2017). 한편 인터넷 미디어의 발달로 인해 개인들도 조세정책에 관해 자유롭게 자신의 의견을 표출할 수 있는 여론형성 여건이 갖추어졌다. 이러한 변화는 정부의 독단적인 정책결정을 견제

---

1) 2014년 귀속 근로소득세 연말정산 대란에 따른 세법수정안 발표가 대표적인 사례이다. 2013년에 8월 8일 발표된 세법개정안은 근로소득자 세 부담 증대에 대한 부정적 여론이 형성되자 일주일도 안 되어 수정안이 발표되었다(국회예산정책처, 2013). 2015년에는 귀속 연말정산을 받은 납세자들의 불만이 가중되자 정부는 세액공제 혜택을 확대하는 내용의 「2015년 연말정산 보완대책」을 발표하고 이내 소급적용하기도 하였다(기획재정부, 2015).

2) 엄밀하게 말하자면 이는 승낙과세원리(‘No Taxation without Representation’)에 기초한 영국식 조세법률주의를 의미한다(황남석, 2016).

하고 조세정책이라는 정치적 과정에서 민의가 반영되고 사회적 합의를 달성할 수 있게 되는 환경이 갖추어진 것으로 평가할 수 있다. 그러나 만일 다수결의 원칙이라는 공정한 민주적 절차를 통해 권한을 위임받은 대의기관의 결의와 정부의 행정활동이 왜곡되거나 편향된 인터넷 미디어 여론의 영향을 받게 된다면 진정한 의미에서 민의가 반영되는 과정이라고 보기는 어렵다. 이러한 측면에서 조세정책에 반영되는 의견이 여론 대표성을 가지고 있는지 평가하는 것은 중요한 의미를 지닌다. 인터넷 여론이 여론을 대표하지 못한다면 국민의 의사를 대표하는 국회의 여야 합의를 통해 국민적 합의에 이른 조세정책이 일부 납세자의 편향된 의견에 의해 긴급히 수정 또는 철회되는 것은 사회적 합의를 번복하는 것으로 공정하지 못하며 세제 안정성을 저해하기 때문이다(경향신문, 2021). 결국 인터넷 미디어 여론이 조세정책에 영향을 미치는 것이 정당한지에 대한 질문은 해당 미디어 여론이 조세정책에 대한 국민의 선호를 얼마나 잘 반영하고 있는가의 질문으로 환원될 수 있다.

황의찬・우석진(2022)이 지적한 바와 같이 인터넷 미디어 여론이 정책수립 과정에 영향을 줄 것이라는 사회적 인지와 그 중요성에도 불구하고, 이러한 여론이 정책과정에 어떠한 영향을 미치는지 또는 어떻게 수렴되고 있는지를 수치화한 연구는 발견하기 어렵다. 이는 인터넷 미디어 여론이 뉴스 댓글이라던가, 청와대 국민 청원, 소셜 미디어 등과 같이 텍스트의 형태로 나타나기 때문에 수치화하기 어려운 성격을 가지고 있기 때문이라고 할 수 있다. 이러한 한계에도 불구하고 최근 발달하고 있는 텍스트 마이닝 기법은 뉴스나 댓글과 같은 텍스트 정보를 수량화된 배열로 나타내어 분석하는 이론적 틀을 제시하여 텍스트를 정량적으로 분석할 수 있는 기반을 제공하고 있다. 본 연구에서는 비정형 데이터인 텍스트 자료를 활용하여 기존에는 정량적 분석이 어려웠던 주제인 인터넷 여론의 조세정책 관련 여론 대표성을 검토하고자 했다는데 의의가 있다.

따라서 본 연구에서는 Gentzkow et al. (2019)의 텍스트 분석 이론을 바탕으로, 인터넷 미디어 여론이 조세정책에 관한 국민의 선호를 얼마나 잘 반영하는지 점검하고 정책 일반에 적용할 수 있는 시사점을 도출하였다. 이를 위해 인터넷 기사 댓글과 국민을 대표할 수 있는 표본을 대상으로 실시한 설문조사 결과의 정량적 비교를 통해 댓글의 여론 대표성을 검증하였다. 댓글과 설문조사에서 각각 추출된 세부주제가 상이하게 나타나거나 찬반의 비율이 지나치게 다르게 나타난다면 댓글의 여론 대표성을 의심해볼 수 있다. 우선 무작위 추출된 납세자 표본에 대해 독립적으

로 진행한 세 차례 설문조사[3]에서 가상자산 기타소득세 부과에 대한 찬반 여부와 해당 이유를 서술한 문장[4]으로 긍정과 부정의 꼬리표(label)가 붙여진 학습데이터를 구성하였다. 이 학습데이터로 사전학습(pre-trained)된 트랜스포머(transformer) 모형을 추가학습(fine-tuning)함으로써 조세정책에 대한 납세자 의견을 긍정/부정으로 분류할 수 있는 감성분류기(sentiment lexicon)를 구축하였다. 인터넷 미디어 여론 데이터로는 2018~2022년간 가상자산 기타소득세 도입 관련된 주요 언론사 인터넷 기사의 댓글을 수집하였다. 학습(train)과 검정(validation)된 감성분류기를 통해 연도별 기사 댓글의 긍정/부정 댓글 비율과 설문조사 결과 얻어진 긍정/부정의 비율을 비교하여 댓글의 편향도(slant)를 점검하였다. 마지막으로 댓글과 설문에서 얻은 의견들에 담긴 주제를 베이지안 추정에 기반한 토픽모형인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)을 통해 각각 추정한 후 주요 주제들의 차이를 살펴보았다.

위와 같은 방법으로 구축된 감성분류기는 학습결과 정확도, 정밀도, 재현율, F-1 score가 모두 0.93의 값을 가지는 양호한 분류성능[5]을 확인할 수 있었다. 이렇게 구축된 감성분류기를 통해 댓글 텍스트의 감성을 분류한 결과 인터넷 미디어 여론은 조세정책에 대해 매우 부정적으로 편향된 것을 확인할 수 있었다. 설문의 경우 가상자산 기타소득세에 대한 긍정의견 비율이 약 69.3%였으나, 댓글의 경우 긍정의견 비율이 최대 16%, 부정 비율은 최소 80% 이상으로 나타났다. 또한 댓글에 대해 독자들이 표시하는 '좋아요'와 '싫어요' 정보를 가중치로 이용하여 분석한 결과 댓글의 부정 강도가 더욱 높아지는 것을 확인할 수 있었다. 이는 조세정책에 대해서 원색적이고 자극적인 부정 댓글들이 긍정적으로 쓴 댓글보다 더 많은 공감대를 형성[6]하며 여론을 유도하고 있는 것을 암시한다. 또한 부정적으로 쓴 댓글이 훨씬

3) 지역, 연령, 교육정도, 소득, 성별 등을 종합적으로 고려하여 세 차례에 걸쳐 독립적으로 실시하여 피설문자로 추출된 표본이 국민을 대표한 수 있도록 설계하였다.

4) 설문내용에는 가상자산 기타소득세 과세에 대한 찬성(긍정)과 반대(부정)를 선택하도록 하였으며 댓글과의 비교가능성을 위해 댓글과 같은 형태로 의견이 표출될 수 있도록 찬성 또는 반대하는 이유를 문장으로 적어내도록 하였다.

5) Lee, Kim and Park(2019a)에서 분류성능지표 0.68로 학습된 감성분류기로 추출한 금융통화위원회 의사록의 논조가 기준금리를 0.1% 유의수준 하에 정확히 예측한다는 점을 감안할 때, 분류성능지표가 0.93라는 것은 상당히 양호한 성능임을 나타낸다.

6) 종종 자극적이고 폭력적인 댓글이 독자의 감정에 호소하며 독자의 공감대를 형성한다. 온라인 글쓰기와 읽기에서 소수 활동가의 지배가 강력하게 나타난다는 김은미・이준웅(2006)의

많은 '좋아요'를 받았다는 것은 관련 조세정책에 대한 인터넷 여론을 조성하는 이들 대부분이 해당 조세정책에 부정적인 의견을 가진 사람들로 편향되었을 가능성을 시사한다. 따라서 정책적 의사결정이 인터넷 여론에 민감하게 반응하여 이미 국민적 합의가 이루어진 정책을 번복하는 것은 공정성을 저해하며 정책의 안정성을 해칠 우려가 있다.

확률적 토픽모형[7]으로 추출한 주제의 일관성 측면에서도 인터넷 댓글의 경우 가상자산소득 과세에 대한 기사의 댓글에 기사의 주제와 무관한 댓글이 달리거나 정책에 대한 선호보다 현 정부나 정당에 대한 지지여부가 나타나는 등 비일관된 주제가 추출되고 있다. 예를 들어, 2018년도의 댓글 44%가 부동산 관련 주제였으며, 2019년도의 댓글 중 35%는 금융투자소득세, 34%는 현 정부에 대한 일반적인 불만 등을 나타내는 주제로 나타났다. 반면 가상자산 기타소득세 관련 설문조사에서 얻어진 텍스트들을 주제 분석을 했을 때는 일관성 있게 해당 조세정책에 대한 주제만으로 구성되어있는 것이 확인된다. 이와 같은 분석결과를 토대로 인터넷 댓글은 주제와 찬반비율의 편향성이 있다고 결론지을 수 있다. 따라서 국민적 합의를 바탕으로 수립된 정책이 긴급히 수정되는 등 정책이 댓글과 같은 인터넷 여론에 민감하게 영향을 받는다면 이는 국민 여론 수렴 과정의 공정성을 저해할 수 있다.

본고는 다음과 같이 구성되어있다. 제Ⅱ장에서는 텍스트 마이닝 방법론을 이용하여 조세정책 또는 여론의 편향성을 연구한 선행연구들을 소개하고 본 연구와의 연관성을 논한다. 제Ⅲ장에서는 최신 개발되어 널리 활용되고 있는 언어모형인 트랜스포머 모형을 활용하여 조세정책에 대한 의견을 나타내는 텍스트의 긍정/부정을 측정할 감성분류기를 구축하는 과정을 상세히 설명하였다. 제Ⅳ장에서는 설문내용을 바탕으로 훈련된 감성분류기로 댓글에 나타난 긍정/부정 비율을 점검하여 설문의 긍정/부정 비율과 비교하고 토픽모형으로 각각의 주제를 추출하여 댓글의 여론 대표성을 평가한다. 제Ⅴ장에서는 연구결과가 시사하는 정책적 함의에 대해 논의한다.

연구의 내용과도 일치한다.

7) 후술하는 바와 같이 베이지안 확률모형에 근거한 Latent Dirichlet Allocation 방법으로 기사, 댓글, 설문내용의 주제를 각각 추출하였다.

Ⅱ. 선행연구

텍스트를 데이터로 사용한 연구는 여러 분야에 폭 넓게 적용되고 있으며 활용되는 텍스트의 종류도 다양하다. 언론의 기사의 경우 보편적으로 구할 수 있는 텍스트 데이터의 원천이므로 경제와 정책 연구에 언론의 기사를 활용하는 경우가 많다. Groseclose and Milyo (2005) 와 Gentzkow and Shapiro (2010) 는 텍스트 마이닝 방법론을 통해 언론사들의 정치적 편향성을 정량적으로 연구하였다. Groseclose and Milyo (2005) 는 미국의 언론사들의 정치적인 성향을 측정하기 위하여 우선 정치인들의 성향을 측정한 후, 이들 정치인들의 연설에서 사용되는 단어들이 각각의 언론사에서 얼마나 자주 활용되는가를 통해 언론사들의 정치적 성향을 정량적으로 측정하였다. 분석결과 언론사들은 정치인들에 비해 중도적인 경향을 가지는 경우가 많았으며 평균적으로는 약간 진보적인 경향이 있는 것으로 확인 되었다. Gentzkokw and Shapiro (2010) 의 경우 Groseclose and Milyo (2005) 의 방법론을 개선하여 미국의 신문사들의 정치적 성향(media slant) 을 측정하였다. 저자들은 연방의회 의사록(Congressional Record) 에 사용된 단어들 중에서 발언자의 소속정당을 가장 잘 예측할 수 있는 단어들을 통계적으로 추출하였다. 이러한 방법론은 Groseclose and Milyo (2005) 에 비해 정치적 성향을 측정하는데 사용되는 단어들의 선택에 있어서 자의성을 크게 감소시켰다.

텍스트에 내재된 어조(tone) 을 측정하여 데이터로 활용하기도 하는데, 이러한 감성분석(sentiment analysis) 은 주관적 의견을 표출한 텍스트 데이터에서 찬성/반대, 긍정/부정 등의 주관적 어조를 머신러닝을 통해 분류하는 작업이다. 감성분석이 활용되던 초기 단계에는 문서에 나타날 수 있는 단어별로 긍정 또는 부정 감성을 사전 할당하고, 해당 단어가 나타나는 빈도 등으로 문서의 긍정/부정 어조를 분류하였다. 감성분석의 초기 연구는 Baker et al. (2016) 가 편제한 EPU (Economic Policy Uncertainty) 가 대표적으로 미국의 10대 일간지 기사에서 'economy', 'policy', 'uncertainty'등 단어가 등장하는 기사의 비율로 일일 불확실성을 측정하는 방식이다. EPU는 문서의 감성을 지나치게 단순하게 측정한다는 일부 비판에도 불구하고 현재까지 가장 널리 활용되는 지표로 자리매김하였다. 이외에도 Nyman et al. (2021) 은 영란은행(Bank of England) 과 기사에 사용된 단어를 흥분(excitement) 과 우려(anxiety) 두 분류로 구분하여 시장 심리를 추출하였다. Lucca and Trebbi

(2011)는 FOMC 의결문에 사용된 단어들을 동반출현확률(Semantic Oriented-Pontwise Mutual Information) 분류법을 사용해 긍정과 부정 부류로 구분하고 의사록의 어조를 추출하였다. 이후 좀 더 정교한 분류기법을 활용한 연구들이 등장하는데 단어의 조합에 긍정/부정을 정의하는 방식으로 단어의 조합이나 문맥에 의하여 달라질 수 있는 감성을 정밀하게 분류하고자 하였다. 이는 ECB 의사록의 논조를 추출한 Picault and Renault(2017)나 금융통화위원회 의사록 논조를 추출한 Lee, Kim and Park(2019a)과 같이 주로 중앙은행의 의사록에 담긴 주관적 견해에 대한 분석에 적용되었다. 이와 같이 비구조적인 텍스트와 언어로 전달되는 정보가 통화정책 향방을 나타내주는 중요한 정보자원으로 연구된 사례가 많다.

한편 논조와 감성 분석을 조세정책에 적용한 연구는 많지 않은데 국내에서는 황의찬・우석진(2022)이 납세자의 정서가 정부의 조세정책에 미치는 영향을 분석한 사례가 유일하다. 황의찬・우석진(2022)은 2013-2020년의 세법개정안 중 조세저항이 있었을 것으로 예상되는 12개의 개정안[8]에 대한 감성지수를 산출하였다. 그리고 이 감성지수가 정책 수정 혹은 철회 확률에 어떠한 영향을 미쳤는지 분석하였다. 분석결과 납세자의 부정적인 정서를 대표하는 감성지수가 1단위 낮아지면 정부가 정책을 철회할 가능성이 약 1.5% 증가한다고 주장하였다[9].

텍스트 마이닝 방법론 중 텍스트를 압축하거나 추정하는 등의 방법으로 주제를 추출하는 기법이 있는데 이를 토픽모형(topic modeling)이라고 한다. Zhao et al.(2011)의 경우 토픽모형인 잠재디리클레할당(Latent Dirichlet Allocation, LDA)을 통해 트위터와 전통적인 언론인 뉴욕타임스가 다루는 뉴스의 주제들이 어떻게 다른지 분석하였다. 저자들은 LDA 모형을 통해 트위터와 뉴욕타임스가 다루는 뉴스들의 주제를 분류하였으며 분류된 주제들을 사건중심(event-oriented), 인물중심(entity-oriented), 일반(long-standing) 카테고리로 분류하였다. 분류 결과 트위터의 경우에는 뉴욕타임스보다 인물중심의 뉴스들이 상대적으로 더 많고 사건 중심의 뉴스들의 빈도는 낮은 것으로 확인되었다. Hansen and McMahon(2016)은 잠재디리클레할당으로 FOMC의 의결문을 현황에 대한 판단과 사전적 정책방향 제시(forward guidance)로 분류하고 각 주제에 대한 논조를 측정하였다.

---

8) 2013년 소득세 공제방식 변경, 2014 담뱃세 인상, 2017년 소득세・법인세 최고세율 인상 등 12개의 조세저항 이벤트(우석진・황의찬, 2022)

9) 통계적으로 유의성이 높지는 않았다.

본 연구는 텍스트 분석 방법론을 사용하여 조세정책 관련 인터넷 미디어 여론이 조세정책에 관한 국민들의 선호를 잘 반영하고 있는지 연구한다. 황의찬·우석진(2022)은 세 부담이 증가되는 조세정책에 대한 부정적인 인터넷 미디어 여론이 정부가 해당 조세정책을 수정하거나 철회하는데 영향을 주는 것을 보였지만, 인터넷 미디어 여론이 조세정책에 대한 국민들의 실제 선호를 잘 반영하고 있는지에 대해서는 논하고 있지 않다. 텍스트 마이닝 방법론은 위의 사례와 같이 여론의 편향성을 연구하는데 자주 활용되어 왔으므로 본 연구의 목적에도 적합할 것으로 보인다.

Ⅲ. 조세정책 수용성을 측정할 감성분류기의 구축

머신러닝 기법의 발달과 함께 데이터를 분류하고 분석하는 기법 또한 발전하였다. 텍스트는 여러 사람들로부터 생산되므로 그 양이 방대할 뿐만 아니라 작성과 동시에 그 의미를 해석할 수 있으므로 텍스트를 활용하여 투표 결과를 예측하거나 시장의 기대를 측정하는데 널리 사용되고 있다(김수현 외, 2020; Lee, Kim, and Park, 2019b). 또한 텍스트를 활용하면 시행이 예정된 정책에 대한 국민의 반응을 즉시 점검할 수도 있을 것이다. 다만 방대한 텍스트로부터 일관된 정보를 추출하기 위해서는 객관적이고 정량적인 분석도구가 필요하다. 따라서 본 절에서는 가상자산 기타소득세 관련 개방형 설문자료를 바탕으로 감성분류기를 구축한다.

1. 감성분류기의 정의와 구축방법

시장의 기대를 측정하고 정책에 대한 반응 등을 점검하는 데는 감성사전(sentiment lexicon) 또는 감성분류기(sentiment classifier)를 필요로 한다. 감성분류기는 학습된 모형을 통해 특정 단어 또는 형태소의 조합이 어떠한 감성을 띄는지 측정함으로써 문장 및 텍스트 전체의 감성을 정량적으로 측정한다. 감성분류기를 구축하기 위해 우선 모형을 학습할 수 있는 자료가 필요한데 해당 자료에는 특정 텍스트(문장, 단어 등)가 지닌 감성이 구체적으로 표기되어있어야 한다. 따라서 학습자료를 얻기 위해 단어 및 형태소 조합의 감성을 사람이 표기하거나 텍스트의 감성을 분류할 수 있는 지표를 통해 표기할 수도 있다.

학습자료가 준비되면 감성분류기의 모형을 선택해야 한다. 모형은 분류와 예측

성능이 주요한 기능이므로 머신러닝을 활용한다. 여기에는 로지스틱 회귀, 서포트 벡터 머신(SVM, Support Vector Machine), 나이브 베이즈 분류기(naive Bayes classifier)와 같이 분류에 특화된 모형을 사용하거나 비선형 모형인 각종 인공신경망(neural networks)을 사용할 수 있다. 특히 인공신경망은 성능과 활용 가능성 등의 이점이 있어 최근 들어 적용범위를 더욱 넓혀가고 있다.

학습자료는 텍스트의 전처리(preprocessing) 과정을 거친다. 전처리 과정은 텍스트를 정제하여 컴퓨터의 알고리즘이 읽을 수 있는 형태로 변환하거나, 정보량이 없는 관용적인 말들을 제거하는 등 텍스트의 질적 보완이 이루어지는 과정이다. 텍스트의 전처리는 토큰화(tokenizing)로 부터 시작하는데 토큰화 과정은 어간과 어미를 분리하는 어간분리(stemming)와 변형된 동사와 형용사를 원형으로 다시 변환하는 원형복귀(lemmatization) 과정을 거친다. 그 이후 숫자, 기호 및 불용어[10] (stopwords)를 제거하여 감성측정에 활용할 수 없는 텍스트를 제거한다. 한글 텍스트의 경우에는 원형복귀보다는 형태소(morpheme) 추출 과정을 거쳐야 하는데 한글에서는 형태소가 의미를 지닌 언어의 최소단위이기 때문이다. 추출된 형태소를 조합하면 감성을 분석할 수 있는 토큰이 완성되며 전처리를 마무리하게 된다.

2. 조세정책 수용성을 측정할 감성분류기 구축

본 연구에서는 개방형 주관식 설문조사를 통해 긍정과 부정 의견이 표기된 학습자료를 구성하였다. 피설문자는 〈Table A-1〉과 같이 지역, 연령, 교육정도, 소득, 성별 등을 종합적으로 고려하여 전체 과세대상자를 대표할 수 있는 표본[11]으로 구성하였다. 총 3회에 걸친 설문조사에서 "정부는 가상자산에서 발생한 소득에 대해 20%의 세율을 적용하여 과세하는 것을 검토하고 있습니다. 귀하는 가상자산 소득에 대해서 과세하는 것에 대해 어떠한 입장이십니까?"라는 질문에 대해 긍정 또는 부정으로 답하도록 하였다. 긍정으로 답변한 경우 "가상자산 소득 과세에 긍정적으로 생각하는 이유는 무엇입니까?"라는 추가질문에 대해 이유를 문장으로 적도록 하

10) 영어의 'the'나 한글의 '이, 그, 저'와 같이 빈번히 사용되나 그 자체로는 감성을 내포하지 않는 단어.

11) 표본은 가상자산에 대한 국민들의 평균적인 인식을 반영하기 위해 가상자산에 대한 지식 유무 등과 무관하게 구성하였다.

였고, 부정으로 답변한 경우에는 "가상자산 소득 과세에 부정적으로 생각하는 이유는 무엇입니까?"라는 설문을 통해 그 이유를 답변하도록 하였다. 위 설문 결과 1회차에서 총 715개(긍정: 479개, 부정: 236개), 2회차에서 총 709개(긍정: 501개, 부정: 208개), 3회차에서 총 708개(긍정: 497개, 부정: 211개)의 답변을 표본으로 구하였다. 3회에 걸친 설문이 독립적으로 추출된 표본에 대해 이루어졌음에도 불구하고 각각의 긍정 답변의 비율은 각각 67.0%, 70.7%, 70.2%으로 긍정 비율이 70% 내외로 나타났다. 이러한 표본의 구성과 설문답변 긍정비율의 일관성으로 볼 때 각 설문조사는 가상자산 소득 과세에 대한 민간의 인식을 잘 반영[12] 하고 있다고 할 수 있다.

(1) 감성분류기: 트랜스포머 모형

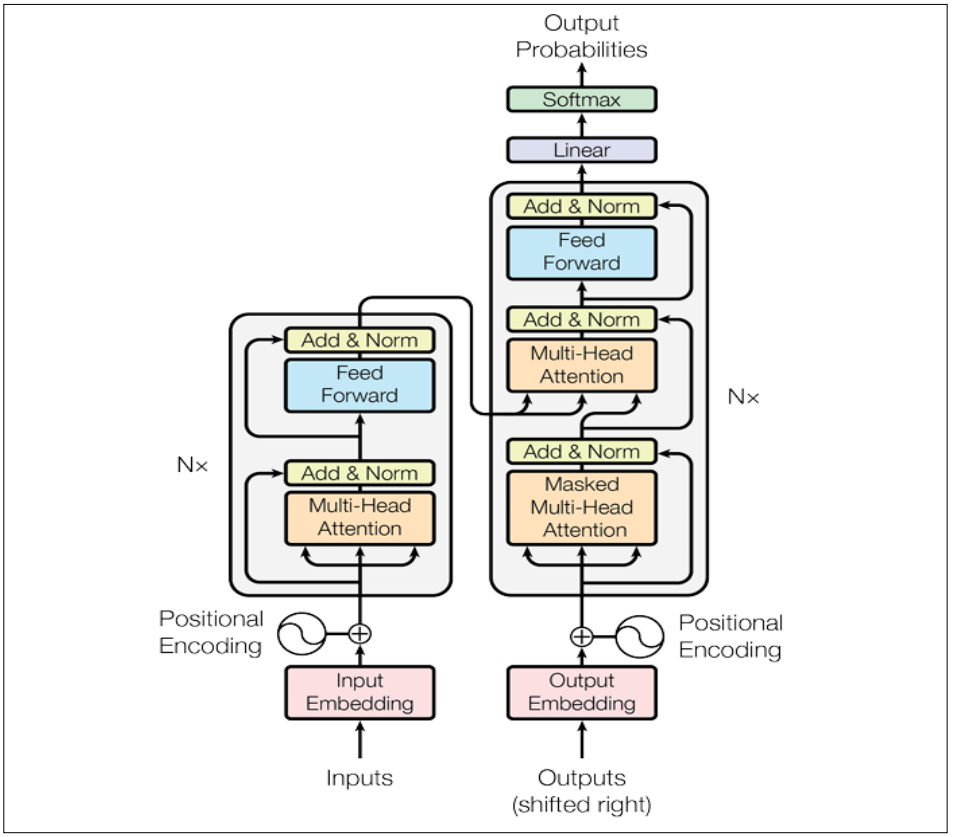
머신러닝과 자연어분석의 발달로 감정분석기를 구축하는 방법은 다양하다. 본 연구에서는 가장 최신에 개발이 되어 자연어처리뿐만 아니라 이미지 처리 등 다양한 분야에서 월등한 성능을 인정받는 트랜스포머(transformer) 모형으로 감성분류기를 학습하였다. 트랜스포머는 2017년 구글(Google)이 제안한 seq2seq 딥러닝(sequence-to-sequence deep learning) 모형으로 기계번역을 위해 개발된 거대 인공지능 모형이다(Vaswani et al., 2017). 기계번역은 입력된 언어(source language)의 단어 배열을 다른 언어(target sequence)의 배열로 변환하는 과정이다. 트랜스포머는 이러한 자동번역(machine translation)을 위해 개발된 모형이나, 기존 번역모형에서는 난해했던 언어 간 서로 다른 배열의 길이(예: 한글 5개 단어를 번역하면 영어로 7개 단어가 되는 경우) 문제에서 자유로울 뿐만 아니라, 속성이 다른 자료 간 변환(예: 문장⇒이미지)도 가능하다는 점이 특징으로 모형의 이름(트랜스포머)이 그 특징을 잘 설명한다. 트랜스포머에서 자료의 변환이 가능한 이유는 입력 자료의 특성을 압축하는 인코더(encoder)와 압축된 특성을 해제하며 원하는 자료의 형태로 출력하는 디코더(decoder)로 구성이 되어있기 때문이다.

트랜스포머의 학습과정은 입력 자료의 압축(encoding)과 해체(decoding)과정에서 정답에 해당하는 배열의 확률을 높이는 과정으로 이루어진다. 일반적으로 트랜

12) 긍부정의 이유로 제시한 문장들도 구체적인 이유를 제시한 답변이 많아 가상자산 소득 과세에 대한 감성을 측정할 감성분류기를 학습하기에 적합하다.

스포머는 압축과 해체를 통해 입력 자료와 다른 형태(예: 다른 언어, 이미지 등)을 출력해내지만, 감성분류기를 구축하기 위해 필요한 출력은 '긍정' 또는 '부정'이다. 따라서 감성분류기 구축을 위해 학습할 디코더 부분은 다른 언어모형에 비해 간단한 구조를 지닌다. 트랜스포머가 기존의 딥러닝 모형과 구분되는 특징은 집중(attention) 또는 자기집중(self-attention) 알고리즘에 있다. 이는 입력 자료 중에 정보량이 많은 자료에 대한 입력가중치를 높이는 방식으로 학습이 이루어진다는 의미이다. 기존 언어모형에서 활용되던 재귀신경망(recurrent neural network, RNN)에서는 단어의 입력 순서에 따라 정보량의 가중치가 달라지는 한계가 있었으며 합성곱 신경망(convolutional neural network, CNN)에서는 사전 정의된 초모수(hyperparameter)인 필터의 크기를 초과하는 문장에 포함된 단어 간 관계를 읽어낼 수 없다는 한계가 있었다. 따라서 재귀신경망을 사용할 경우 문장에서 가장 중요한 단어가 앞쪽에 나타날 경우 모형의 성능 저하로 이어질 수 있으며 합성곱신경망의 경우 문장이 길어질 경우 모형이 문장의 문맥을 분석해내지 못할 가능성이 있다. 자기집중 알고리즘으로 학습하는 트랜스포머 모형의 경우 전체 문장에서 밀접한 관계를 지닌 단어 조합의 모든 경우의 수에 대해 정보량을 측정하여 높은 정보량에 해당하는 조합에 집중함으로써 모형의 성능을 개선한다. 감성분류기에서도 단어 간 관계를 중점적으로 고려하므로 분류 성능을 높일 수 있다. 예를 들어 "가상자산 소득 과세에 긍정적으로 생각하는 이유는 무엇입니까?"라는 설문에 대해 "가상자산 투자는 안전하지 않으므로 위험한 투기행위를 방지하기 위해"라고 답변한 문장과, "가상자산은 개인이 위험을 감수하고 투자하므로 정부가 가상자산 거래시스템의 안전을 보장하지 않는 이상 과세하는 것은 이치에 맞지 않음"이라고 부정의 이유를 답변한 경우가 있다고 가정하자. 두 답변 모두 {가상자산, 위험, 안전}이라는 단어의 조합이 들어있음을 알 수 있다. 그럼에도 불구하고 첫 번째 답변은 '긍정'으로 분류되어야 하며 두 번째 답변은 '부정'으로 분류되어야 한다. 따라서 같은 단어조합이 있는 서로 다른 문장을 각각의 감성으로 분류해내기 위해 {가상자산, 위험, 안전}과 여타 단어들의 조합이 문맥을 이룸으로써 각각 긍정과 부정의 감성으로 분류되도록 학습되어야 한다.

〈Figure 1〉 The Schematic Structure of the Transformer Model



Note: The schematic structure is a general representation of the Transformer model, and the sentiment classifier in this paper has a simple structure with only positive and negative outputs.

Source: Vaswani et al. (2017).

(2) 감성분류기 학습

본 연구에서 구축한 감성분류기는 구글이 Vaswani et al. (2017)의 이론을 기초로 개발하여 배포한 BERT(Bidirectional Encoder Representations from Transformers)를 응용한 것[13]이다. BERT(Devblin et al., 2018)는 위키피디아(Wikipedia)의 25억 개 단어와 BookCorpus[14]의 8억 개 단어로 사전학습(pre-trained)된 언어모형이다.

13) BERT를 한글에서 구현하기 위해 많은 library가 개발되어있다. 본 연구에서는 eKorpKit (Lee, 2022)을 사용하였다.

사전학습된 모형이므로 감성이 표기(labelling)된 학습자료로 추가학습(fine-tuning)할 경우 높은 성능을 나타내는 것으로 알려져 있다. BERT의 기본구조는 트랜스포머(12개 또는 24개)를 직렬층(sequential layers)으로 중층 구성하였으며 단어 간의 관계인 문맥을 인지하는 트랜스포머 모형의 특성이 잘 구현되어있다. 그러나 감성분석은 분야(domain)에 따라 분류성능이 크게 좌우되므로 분야별로 추가학습한 다양한 BERT가 공개되었다. 예를 들어 1,800백만 건의 의학, 생명학 논문으로 추가학습한 의학과 약학 분야의 BioBERT가 대표적이다. 금융, 경제분야에서는 약 20만건의 미국 상장기업의 연차(10-K) 및 분기(10-Q) 보고서[15], 약 14만건의 기업실적보고회 원고(earnings call transcript), 약 49만 건의 S&P 기업분석 보고서 등에서 얻은 49억개의 단어로 학습한 FinBERT가 있다. FinBERT의 경우 사전학습만으로 구성된 BERT에 비해 최대 16.1%의 성능향상을 보여준다. 본 연구에서는 한글 위키피디아 자료로 사전훈련된 KR-BERT를 설문의 내용으로 추가학습[16]하여 감성분석기를 구축하였다.

가. 학습자료의 전처리

학습자료는 3회에 걸친 개방형 설문 조사 조사이며 설문내용은 감성분류기의 학습을 염두에 두고 설계하였으므로 긍정과 부정의 이유를 문장으로 답변하도록 하였다. 답변은 대체적으로 성실하게 이루어졌으나 일부 불성실한 답변의 경우 감성분류기 학습결과의 편의(bias)를 야기하는 원인이 되므로 〈Table A-2〉 및 〈Table A-3〉와 같이 원표본에서 제외[17]하여 수정표본을 구성하였다.

---

14) BookCorpus는 무료배포된 11,038권 소설의 약 7천4백만 문장으로 구성된 거대 텍스트 자료이다.

15) 미국 증권감독위원회(SEC)에 공개된 상장기업의 의무보고서로 기업의 영업 및 재무상태가 상세히 기술되어있다. FinBERT에서는 영업활동(Item 1), 위험요소(Item 1A), 경영 및 분석(Item 7) 섹션을 활용하였다고 되어있다.

16) 설문의 내용은 비교적 정형적인 문장과 단어로 구성되어 검열을 피해 변형된 단어들이 즐비한 인터넷 댓글과는 단어와 문장구조 측면에서 차이가 있다. 그럼에도 불구하고 검열을 피해 변형된 단어들은 주로 욕설에 해당하며 이들은 대부분 부정적인 감성을 띄게 되므로 설문의 내용으로만 학습한 감성분석기를 사용한 경우에도 본 연구의 분석결과와 결론에는 영향이 없다.

17) 이는 모형추정의 편의를 야기하는 이상치(outlier)를 제거하는 winsorizing과 같은 맥락이다. 부록의 〈Table A2〉에는 설문조사결과 가상자산 과세에 대한 긍정의 의견을 밝힌 피설문자가

설문 자료를 활용하여 모형을 학습하는데 다른 장애요인은 긍정과 부정의 비율이 큰 폭의 차이를 보이는 불균형(unbalanced) 자료라는 점이다. 학습자료의 긍정과 부정 비율이 큰 폭으로 차이를 보일 경우 비율이 높은 쪽으로만 분류하더라도 최소한 그 비율만큼의 정확도를 나타내므로 학습이 정상적으로 진행되지 않을 가능성이 매우 높다. 3차에 걸친 설문자료가 모두 70% 수준의 긍정비율이 나타나므로 감성분류기가 모든 문장을 긍정으로만 분류해도 최소 70%의 정확도를 갖게 된다. 이 경우 부정의 문장도 긍정으로 분류함으로써 실질적 분류를 할 수 없는 감성분류기가 학습된다. 따라서 자료를 균형 있는 자료로 보완해야할 필요가 있는데 본고에서는 무작위 배치추출(random batch sampling) 방법[18]으로 부정의 문장에서 중복 추출하여 표본의 수를 긍정 문장 수와 일치시켰다. 본 연구에서는 전체 답변에서 30%에 불과한 부정 답변에서 중복이 허용되는 무작위 추출 방법으로 긍정과 부정의 비율이 50% 대 50%가 되는 학습자료를 구성하였다.

위와 같이 마련된 텍스트 자료는 어간추출(stemming), 표제어추출(lemmatization), 불용어(stopwords) 제거 등의 전처리 과정을 거쳐 토큰화(tokenizing)하고 품사(part-of-speech, POS) 꼬리표 달기(tagging)와 형태소 분석(morpheme analyzing)을 통해 분석을 위한 자료준비를 마무리한다. 어간추출 과정에서는 단어의 접두어, 접미어, 어미 등을 제거하고 의미를 내포한 어간만 추출하며 표제어추출에서는 변형된 동사, 형용사, 부사를 원형으로 복원시킨다. 불용어 제거 시에는 단어자체만으로는 의미를 내포하지는 않으나 상용적으로 쓰이는 말(이, 그, 저, the 등)을 자료에서 제거한다. 품사 표기는 감성분석에서 주로 활용되는 명사, 동사, 형용사, 부사, 부정어를 추출해내기 위해 반드시 필요한 작업이다. 이후에는 한글의 특성인 형태소 추출과정을 거쳐야 한다. 이를 위한 기본 형태소 분석기는 Python의 mecab

---

긍정의 근거로 제시한 답변 중 실질적인 근거가 될 수 없는 답변을 열거하였다. 이들은 대부분 "이유가 없다", "모른다"등 구체적인 이유가 명시되어있지 않거나 가상자산 과세 찬반 이유와 관련 없는 답변들이 대부분이다. 〈Table A3〉에서는 같은 이유로 학습자료에서 제외된 부정에 대한 근거들을 나열하였다. 부정 의견에서 제외된 답변은 긍정에 비해 수가 적지만 "탈세", "가상"등 긍정 또는 부정의 의견을 뒷받침하지 못하는 단어만 제시한 경우와 "더 올려야 한다"와 같이 오히려 긍정의 의견을 제시한 경우로 학습자료에서 제외하였다. 이들을 제거함으로써 학습된 모형의 정확도(precision) 등을 향상시킬 수 있다.

18) 통계학과 계량경제학에서 추정치의 통계적 검증을 위해 활용되는 bootstrapping 방법과 유사하다. 머신러닝에서도 무작위 배치추출을 통해 샘플의 크기를 늘리는 것만으로도 과적합의 경우를 감소시키고 학습결과가 개선되는 경우가 많다.

library[19] 를 사용하였다.

나. 감성분류기의 학습결과와 성능의 검정(validation)

머신러닝 또는 딥러닝으로 학습한 모형은 모형의 복잡성과 비선형성으로 인하여 이론적으로 정립된 검정통계량으로는 검정이 어렵다(Buckman and Joseph, 2022). 따라서 학습된 모형을 평가하기 위해 정확성, 정밀도 등 예측 및 분류 성능을 평가하는 지표를 산정하여 비교한다. 우선 성능을 평가하기 위한 지표를 산정하기 위해 아래 〈Figure 2〉와 같이 모형의 분류결과와 참값을 비교하는 혼동행렬(confusion matrix)를 구성하고 분류한 결과의 참과 거짓 비율 등을 계산한다.

〈Figure 2〉 Confusion Matrix

Actual \ Predicted	Positve	Negative
Positve	True Positive	False Negative
Negative	False Positive	True Negative

Note: False positive refers to classifying a negative sample as positive, while false negative refers to classifying a positive sample as negative, both of which represent classification errors.

혼동행렬부터 측정할 수 있는 평가지표는 아래와 같이 정확도, 정밀도, 재현율, F-1 score[20] 가 있다. 정확도(accuracy)는 분류의 정확도를 측정하는 측도로 총 표본 수에서 참긍정(true positive) 또는 참부정(true negative)의 비율로 혼동행렬을

19) mecab은 애초에 일본어 형태소 분석을 위해 개발되었으나 한글 형태소 분석을 위해 개조되어 현재까지 개발된 한글 형태소분석기에서 가장 개선된 성능을 보여주고 있다. 본 연구를 위해 사용한 ekorpkit library(Lee, 2022)에도 탑재되어 있다.

20) $\text{정확도(Accuracy)} = \frac{\text{참긍정}+\text{참부정}}{\text{참긍정}+\text{거짓긍정}+\text{참부정}+\text{거짓부정}}$,

$\text{정밀도(Precision)} = \frac{\text{참긍정}}{\text{참긍정}+\text{거짓긍정}}$, $\text{재현율(Recall)} = \frac{\text{참긍정}}{\text{참긍정}+\text{거짓부정}}$,

$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$

활용하여 측정한다. 정밀도(precision)는 감성분류기가 긍정으로 분류한 표본 중 참 긍정을 긍정으로 옳게 분류한 비율이다. 재현율(recall)의 경우 감성분류기가 참긍정 표본 중 긍정으로 옳게 분류한 비율을 나타낸다. F-1 score의 경우 정밀도와 재현율을 하나의 지표로 나타낸 점수이다.

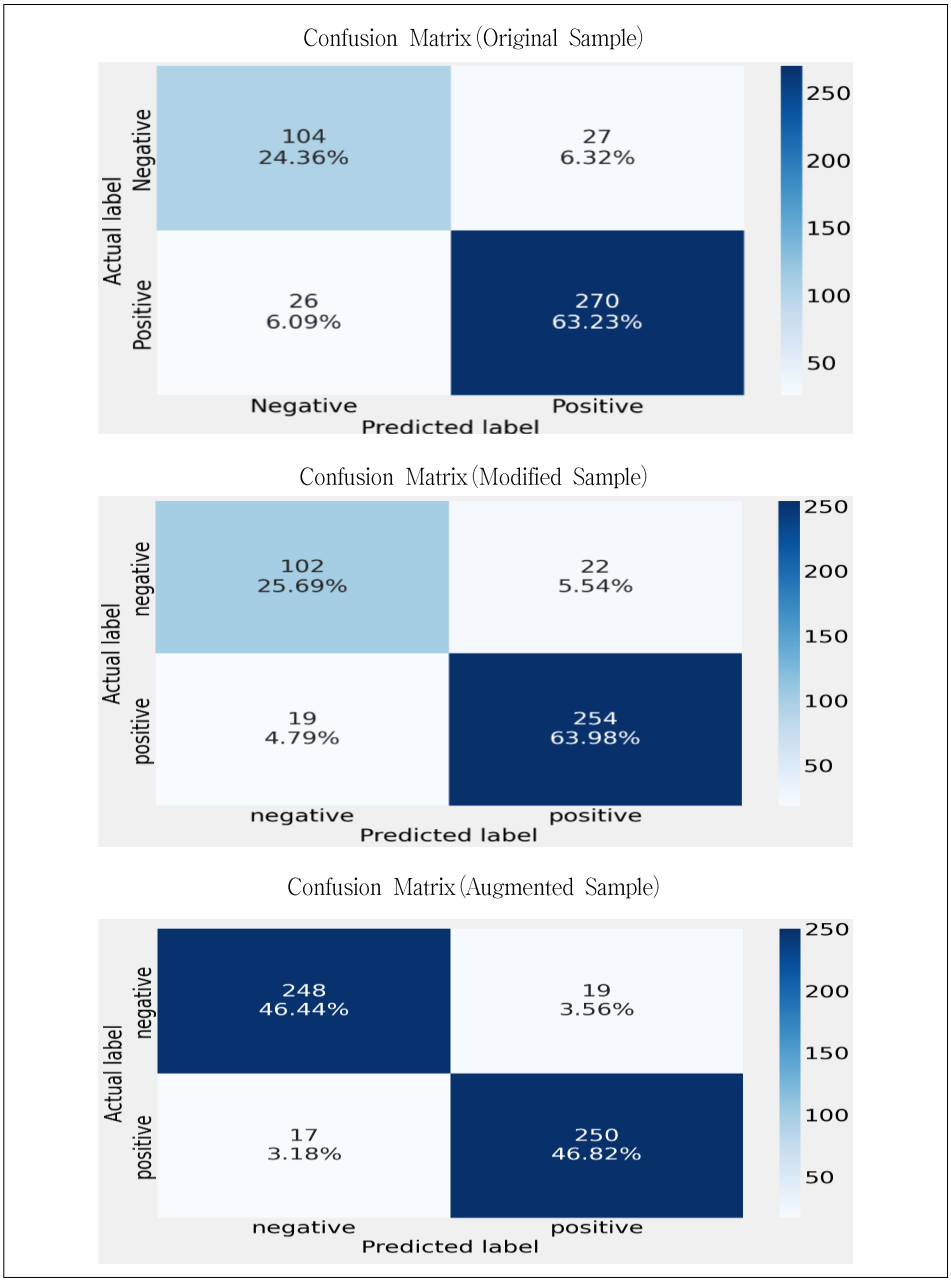
앞서 기술한 바와 같이 감성분류기를 학습하기 위한 최종표본은 무작위 중복 추출을 통해 긍정과 부정의 답변 수를 같도록 만든 표본을 사용하였다. 이를 증강표본이라고 이름하였으며 설문결과를 그대로 사용하는 원표본과 모순되거나 의미가 없는 답변들을 제외한 수정표본과의 학습결과를 비교해볼 수 있다. 〈Table 1〉과 〈Figure 3〉은 각 표본의 수와 학습결과를 비교한 것이다. 원표본의 경우 긍정 1,477개 부정 655개로 긍정:부정이 69.3:30.7 비율이며 수정표본의 경우 68.9:31.1로 두 표본 모두 긍정비율이 70%에 가깝다. 원표본의 학습 결과 평가지표는 정확도, 정밀도, 재현율, F-1 score가 각각 0.86, 0.87, 0.86, 0.86으로 나타났고, 수정표본의 경우에도 네 평가지표가 모두 0.89로 나타났다. 수정표본에서

〈Table 1〉 Sample Sizes and Classification Performance Indexes

	Original sample	Modified sample	Augmented sample
Number of responses			
Positive	1,477	1,364	1,364
Negative	655	617	1,364
Total	2,132	1,981	2,728
train set(70%)	1,492	1,387	1,910
validation set(10%)	213	198	273
test set(20%)	427	396	545
Classification Performance Indexes			
Accuracy	0.86	0.89	0.93
Precision	0.87	0.89	0.93
Recall	0.86	0.89	0.93
F-1 score	0.86	0.89	0.93

Note: The original sample is composed of responses from open-ended survey questions, and the modified sample is a sample that excludes reponses that are inconsistent with answering positively or negatively, or meaningless descriptions. The augmented sample refers to a sample that matches the number of positive and negative responses by randomly duplicating negative responses.

〈Figure 3〉 Confusion Matrix across Different Samples



Note: The classification results are shown according to the confusion matrix structure of Figure 2, and the numbers in the table represent the number of classified samples, while the percentages (%) below the numbers represent the ratio of the classified samples to the total samples.

샘플링을 통해 긍정과 부정 비율을 각각 50%로 맞춘 증강표본의 경우 네 지표가 모두 0.93으로 나타나 앞서 논의한 두 표본에 비해 성능이 크게 개선[21] 되었음을 알 수 있다.

Ⅳ. 인터넷 여론의 감성 측정

앞에서 국민 조세 인식 조사 자료를 활용하여 트랜스포머 모형으로 구축한 감성 분류기로 가상자산소득 과세 관련 뉴스의 댓글의 감성을 분석하고, 분석결과를 국민 조세 인식 조사 자료와 비교함으로써 뉴스의 댓글이 국민의 평균적인 조세 감정을 대표하는지 검증한다. 감성분류기로 기사 댓글의 감성을 분석하기에 앞서 감성 분류기가 사전분류된 댓글의 긍/부정 감성을 얼마나 잘 분류하는지를 점검한다. 만약 감성분류기가 댓글의 감성을 비교적 잘 측정하고 있다면 오차 범위 내에서 댓글의 긍/부정 비율이 3회에 걸친 국민 조세 인식 조사 자료의 결과와 근접하는지 여부를 점검함으로써 인터넷 여론이 국민의 조세감정을 잘 대변하고 있는지를 검증해 볼 수 있다.

1. 댓글의 수집

인터넷 여론 자료는 네이버에서 가상자산 과세 관련 키워드를 기준으로 검색한 뉴스 기사[22] 에 독자들이 쓴 댓글을 수집(crawling) 하였다. 뉴스 기사 자체는 과세에 대해 중립적이고 객관적인 사실만을 전달하므로 감성분석을 위한 자료로는 적합하지 않다. 반면 댓글들은 관련 기사에 대한 독자들의 의견을 표현하므로 기사와 관련된 주제에 대한 일부 여론을 반영한다고 볼 수 있고 감성분석을 위한 자료로도 적합하다. 한편 댓글의 감성을 분석하기 위해서는 댓글이 달려있는 기사들이 가상

21) 일반적으로 텍스트 자료의 경우 측정오차로 인해 네 지표가 1의 값에 가까워질수록 분류기의 성능의 개선이 점차 어려워진다. 그럼에도 불구하고 증강표본으로 학습한 결과 앞선 두 표본의 학습결과에 비해 지표 값이 0.04 이상 상승하는 모습을 보여주므로 성능이 대폭 개선되었다고 할 수 있다.

22) 기계적으로 검색한 기사는 가상자산 과세와 밀접한 관련이 있을 수도 있으나, 암호화폐 과세가 잠시 언급되는 정도로만으로도 해당 기사가 검색되므로 기사를 직접 읽고 필터링하는 과정을 거쳤다.

자산 과세와 밀접히 관련되어야 한다는 조건이 선행되어야 한다. 기사의 주된 내용이 가산자산 과세에 대한 것이라면 댓글은 가상자산 과세에 대한 독자들의 반응을 텍스트로 나타낸 것이라 볼 수 있기 때문이다. 수집된 자료는 플랫폼 서비스인 네이버[23]에서 '과세'를 반드시 포함시켜야할 키워드로 하여 '암호화폐, 가상화폐, 가상자산'을 관련 키워드로, '규제, 종부세'는 제외[24]해야 할 키워드로 하여 뉴스 기사를 검색하고, 검색된 기사와 해당 기사에 독자들이 적은 댓글을 수집하였다.

가상자산 과세에 관련된 기사는 2018년 이후부터 직접 연관된 기사가 검색되므로 기간은 2018.1~2022.6으로 하였다. 자료의 연도별 구분은 감성분류기 구축과는 무관하지만 본고에서는 편의상 연도별로 주제를 점검하였다. 가상자산 과세와 관련된 기사의 수는 2021년까지 지속 증가하다가 2022년 들어 다시 감소하고 있다. 이에 관련 평균 댓글 수는 가상자산 과세 관련 기사가 본격적으로 등장하기 시작한 2018년이 62.3개로 가장 많았으며 2019, 2020년 들어 각각 42.6개, 28.8개로 점차 감소하다가 2021년 들어 다시 53.9개로 급증하였다. 기사에 대한 평균 댓글 수가 기사에 대한 독자들의 관심을 반영한다고 가정하면 가상자산 과세 관련 기사가 등장하기 시작한 2018년에 관심이 가장 높았고 이후 과세에 대한 논의가 지연되었던 2019, 2020년 동안 관심이 잦아들다가 2021년 다시 주목받기 시작한 것으로 해석할 수 있다.

---

23) 특정 플랫폼의 사용자가 국민의 선호를 잘 대변할 수 있는지에 대한 비판이 제기될 수 있으나 해당 플랫폼이 국내 최대 규모의 디지털 뉴스 유통 플랫폼이며 2022년 기준 2644만 명의 구독자 수가 있다는 점(한겨레, 2022)에서 대표성의 문제는 크지 않을 것으로 판단한다. 또한 배진수 외(2022)의 설문조사에 의하면 가상자산 과세에 대한 찬반 응답에 있어서 응답자들의 연령이나 정치적 성향 등은 유의수준 5%에서 통계적으로 유의한 영향이 없는 것으로 나타났으므로 특정 플랫폼의 사용자들이 설령 정치적 편향이 있다고 가정하더라도 가상자산 과세에 긍·부정을 연구하는 본 연구의 결과에는 큰 영향이 없을 것으로 판단한다. 한편 대형 포털의 경우 자체적인 심의 규정과 정부의 규제로 인해 정치적인 편향성을 가지게 될 가능성은 낮은 것으로 평가받고 있으며(임원혁 외, 2019, 210p) 최동욱(2017)은 포털 뉴스 자체의 정치적 양극화 모습은 나타나지 않는다고 보고했다.

24) 플랫폼은 "or" 알고리즘을 바탕으로 검색하므로 여러 키워드로 검색할 경우 각 키워드에 관련된 기사가 모두 검색된다. 따라서 "과세" 키워드와 관련된 기사로 "종부세" 등에 관련된 기사가 있을 수 있으며, "암호화폐"에 관련된 "암호화폐 규제"에 관련된 기사들도 함께 검색된다. 이러한 기사들이 학습자료에 포함될 경우 독자들이 표현하는 감성은 "암호화폐 과세"에 대한 감성과 무관한 댓글들이 많이 포함되므로 자료의 편의(sample bias)를 줄이기 위해서라도 자료에서 미리 배제하는 것이 좋다.

〈Table 2〉 The number of articles and comments related to the taxation of virtual assets by year

	2018	2019	2020	2021	2022
Articles	32	54	120	545	75
Comments	1,995	2,287	3,516	29,629	2,277
Average Comments per Article	62.3	42.6	28.8	53.9	30.4
Number of News Agencies	14	24	31	54	27

2. 기사 댓글을 활용한 감성분류기의 검증

수집한 댓글은 앞서 분석한 개방형 설문조사 자료와 같이 전처리를 통해 감성분석 자료[25]로 준비하였다. 댓글의 감성을 분석하기에 앞서 설문 자료로 학습한 감성분석기가 댓글의 감성을 잘 분석하고 있는지 검증할 필요가 있다. 학습 자료와 검증 자료의 성격이 크게 다를 경우 감성분석기의 정확도 등 평가지표도 하락할 수 있기 때문이다.

검증방법은 수집한 댓글 중 임의 추출한 댓글을 직접 읽고 123개 댓글[26]에 대해 긍정/부정을 분류한 후 동일한 댓글을 구축된 감성분류기로 분류하였다. 분류 결과

〈Table 3〉 Confusion Matrix for Article Comments Classification

Actual \ Predicted	Positive	Negative
Positive	14 (48.3%)	15 (51.7%)
Negative	4 (4.3%)	90 (95.7%)

Note: The classification results are shown according to the confusion matrix structure of Figure 2, and the numbers in the table represent the number of classified samples, while the percentages (%) below the numbers represent the ratio of the classified samples to the total samples.

25) 가상자산 과세 관련 기사는 해당 주제에 대한 객관적인 사실을 서술할 뿐 찬반의 논의를 하지 않는다. 따라서 감성분석은 댓글을 대상으로 하는 것이 적절하다.

26) 분류한 댓글의 수는 1,258개이나 1,135개 댓글의 감성이 "중립"으로 분류됨에 따라 긍정/부정 분류 성능 검증에 사용된 댓글의 수는 123개이다.

〈Table 4〉 Positive Comments Misclassified as Negative

	Comments	Classification probability		Results
		P	N	
Comments contains words that are considered vulgar or have an aggressive sentiment				
1	소득이 있으면 세금을 내야지. 씨밤 여기가 북한이냐??	0.23	0.77	N
2	세금70프로 때려	0.04	0.96	N
3	코인과세 당근해야지 조뺑이치며 돈버는 서민들은 세금 따박따박 내는데 클릭질 몇번해서 세금도 안내고 돈벌게?	0.04	0.96	N
4	그니까 암호화폐 하지 말고 주식하라고... 암호화폐 소득에 세금 혜택 주는 건 미친 짓이지	0.03	0.97	N
5	코인은 뭐하는데 세금혜택을 준다는 것이냐? 악덕 코인들 당장 없애라. … 노동자와 서민들이 느들 봉이냐? 우리들한테 걷은 세금으로 별 그지 같은 사기에 세금혜택을 부여해 달라고 하고 있어. 가면 있으니 누굴 봉으로 아나. 만국의 노동자들이여, 만국의 월급쟁이들이여, 만국의 서민들이여 단결하여 저 간악한 코인 투자자들, 코인 발행자들에게 죽창을 날리자.	0.04	0.96	N
6	비과세를 왜함 과세는 평등해야하지 표얻으려고 쌩 난리 피우지말고 싹 다 과세해라	0.11	0.89	N
Positive expressions conveyed through double negation				
7	비트코인 뭐라고 세금을 안 내는거야	0.03	0.97	N
8	대다수의 국민은 민주당 응원하고, 소수를 위한 과세연기 반대한다	0.08	0.92	N
9	요게 맞지. 주식이야 기업에 투자 개념이지만 코인은 사실상 투기인데 이런 자산에 비과세 혜택을 주는건 이상한거지	0.04	0.96	N
10	만약 코인과세를 하지 않으면...코인은 제도권으로 들어오지 않고, 피해자들이 생긴다. 그리고, 세금을 월급쟁이들이 다낸다	0.04	0.96	N
11	대다수의 국민은 민주당 응원하고, 소수를 위한 과세연기 반대한다	0.08	0.92	N
When sentiment cannot be determined by specific word combinations, but is based on the context				
12	민주당 말 함 들어주고 과세가세요 솔직히 코인도 엄연한 투자잖아요 손해를 감수하고 이득본거에 대한 세금내라고 하는건데...	0.04	0.96	N
13	코인은 무법지대의 주식임. 빨리 코인 자체를 도박 및 불법화 하던가, 아니면 탈중앙화 버리고 세금 메겨서 법제화 하던가.	0.06	0.94	N
Simple error				
14	세금걷자	0.18	0.82	N

Note: There were 15 cases where positive comments were misclassified as negative. This includes duplicated comments in the same article (comment number 12).

는 〈Table 3〉의 혼동행렬과 같다. 우선 123개 댓글 중 부정 댓글 94개 중 90개(95.7%)를 부정으로 분류하였다. 반면 긍정 댓글 29개 중 14개(48.3%)만을 긍정으로 분류하였다. 설문자료로 학습한 감성분류기로 댓글을 분류할 경우 부정 댓글은 높은 정밀도로 분류하는 반면 긍정 댓글의 경우 긍정으로 분류할 수 있는 댓글은 전체의 절반 수준이었다.

감성분류기가 인터넷 댓글의 긍정 댓글의 절반을 부정으로 분류[27] 한데는 〈Table 4〉와 같이 세 가지 이유로 분석해볼 수 있다 첫째, 욕설 또는 비속어와 같이 강한 부정적 감성을 지닌 단어가 쓰이고 있는 경우이다. 해당 단어들에 상대적으로 많은 가중치가 부여되는 경우 문장 전체가 부정감성으로 분류되는 경우가 발생하는 것이다. 다음으로는 부정에 대한 부정으로 긍정을 표현한 댓글의 경우이다. 부정의 부정 문제는 감성분석에서 흔히 거론되는 이슈로, 두 부정 중 하나만 부정으로 인식될 경우에는 긍정이 부정으로 인식되는 오류가 발생한다. 마지막으로 문장 내에 긍정/부정을 나타내는 단어는 없지만 문맥으로만 긍정/부정을 분류할 수 있는 경우이다. 흔히 말하듯 "강한 부정은 강한 긍정"이라는 말과 같이[28] 오로지 문맥으로만 긍정/부정을 판단해야하는 문장이다. 감성분류기의 경우 이 부분에 대한 학습이 필요하나, 문맥만으로 긍정/부정을 판별할 정도의 학습을 진행시킬 경우 오히려 과적합(over-fitting)의 문제가 발생할 우려가 높다. 따라서 부정으로 분류된 긍정댓글을 수정하기 위한 추가 학습과정이나 미세조정 등을 거치지 않는 것이 모형의 일치성(consistency)을 위해서는 바람직할 수 있다. 아래에서는 인터넷 댓글의 감성 분류 시 감성분류기로 분류된 긍정댓글 비율의 두 배 비율까지도 긍정적인 댓글이 있을 수 있다는 점을 감안하여 분석할 것이다.

3. 댓글의 감성분류과 비교 분석

앞서 기술한 바와 같이 3회에 걸쳐 독립적으로 시행된 개방형 설문 조사에서 가상 화폐 과세에 대한 긍정 답변 비율은 68.9%에 이르며 이 비율이 일관되게 나타

27) 부정댓글을 긍정으로 분류한 경우는 부록 〈Table A-4〉에 정리되어있다.

28) 예를 들어 단풍이 든 산을 보며 "이야 경치 죽이는데~!"라고 하면 사람은 이를 매우 긍정적으로 인식하지만 감성분류기의 경우 "죽이다"라는 말에 많은 가중치가 더해지며 부정으로 분류되기 쉽다.

난다. 그렇다면 만약 뉴스에 대한 댓글들이 국민들의 평균적인 조세 인식을 잘 반영한다면 뉴스에 대한 댓글의 감성을 측정할 경우에도 유사한 비율이 나타날 것을 기대할 수 있다. 따라서 위 검증결과를 고려하여 전체 댓글을 긍정과 부정으로 분류하였을 때 국민 조세 인식 조사의 결과와 유사한지 여부를 비교하면 뉴스에 대한 댓글이 국민의 평균적인 여론을 잘 반영하는지 여부를 확인할 수 있다.

댓글은 시간에 따른 과세에 대한 인식 변화를 고려하여 연도별로 구분하여 분류하였으며 결과는 아래 〈Table 5〉와 같다. 2018년부터 2022.6월까지 긍정으로 분류된 댓글의 비율은 각각 0.09, 0.08, 0.08, 0.08, 0.07로 평균 8% 댓글만 긍정으로 분류되었다. 본 연구에서 구축된 감성분류기가 약 50% 확률로 긍정댓글을 긍정으로 분류한다는 위의 검증결과를 반영한다고 하더라도 전체 댓글의 16%만이 긍정댓글이라고 평가할 수 있다. 반면 부정댓글의 경우는 5년간 전체 댓글 중 90% 이상이었다. 검증결과를 고려하여 조정할 경우에도 80% 이상의 비율이다.

댓글의 긍정과 부정의 비율은 댓글에 대해 독자들이 표시하는 '좋아요'와 '싫어요'를 가중치로 하여 측정하였을 때 더욱 큰 폭의 차이를 보여준다. '좋아요'와 '싫어요'는 댓글에 동의하거나 동의하지 않는 독자의 수를 나타내므로 '좋아요' 숫자는 댓글로 작성되지는 않았지만 댓글과 동일한 의견의 개수로 간주할 수 있고 '싫어요'는 반대의견의 갯수로 간주할 수 있다. 따라서 $\mathrm{Like_i}$와 $\mathrm{Dislike_i}$를 각각 i댓글에 대한 '좋아요' 개수와 '싫어요' 개수라고 한다면 $\mathrm{Like_i} - \mathrm{Dislike_i}$를 i댓글에 대한 가중치로 활용할 수 있다. 이와 같이 가중치를 고려한 긍정/부정 댓글의 수는 N_p, N_n, N을 각각 긍정 댓글 수, 부정 댓글 수, 총 댓글 수라고 할 때 아래와 같이 구할 수 있다.

$$\mathrm{Weighted\ Positive} = \mathrm{N_p} + \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{I_p(Comment_i)} \times (\mathrm{Like_i} - \mathrm{Dislike_i})$$

$$\mathrm{Weighted\ Negative} = \mathrm{N_n} + \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{I_n(Comment_i)} \times (\mathrm{Like_i} - \mathrm{Dislike_i})$$

여기에서 $\mathrm{I_p(Comment)} = 1$ if $\mathrm{Comment} \in \mathrm{Positive\ Comments}$,

$\mathrm{I_n(Comment)} = 1$ if $\mathrm{Comment} \in \mathrm{Negative\ Comments}$

위와 같이 가중치를 고려할 경우 긍정과 부정의 댓글 비중은 각각 2018~2022년

평균 4%, 96%가 되어 앞선 비교 결과보다 더욱 큰 폭의 긍정/부정 비율의 차이를 보이고 있다. 이는 긍정적으로 쓴 댓글보다 부정적으로 쓴 댓글이 훨씬 많은 수의 '좋아요'를 받았다는 것인데, 다소 감정적이고 자극적인 부정 댓글일수록 공감대를 형성[29] 하고 있음을 알 수 있다. 이는 인터넷 여론에 기반한 정책 피드백이 국민의 평균적 조세 인식을 왜곡되게 반영할 수 있다는 사실을 입증한다.

〈Table 5〉 Sentiment Classification Results of Comments by Year

(unit: counts)

	Unweighted			Weighted		
	Positive	Negative	Total	Positive	Negative	Total
2018	171 (0.09)	1,826 (0.91)	1,997 (1.00)	1,587 (0.04)	37,060 (0.96)	38,647 (1.00)
2019	186 (0.08)	2,101 (0.92)	2,287 (1.00)	669 (0.04)	18,417 (0.96)	19,086 (1.00)
2020	275 (0.08)	3,241 (0.92)	3,516 (1.00)	680 (0.03)	24,804 (0.97)	25,484 (1.00)
2021	2,404 (0.08)	27,225 (0.92)	29,629 (1.00)	7,489 (0.04)	174,932 (0.96)	182,421 (1.00)
2022	166 (0.07)	2,184 (0.93)	2,350 (1.00)	680 (0.05)	12,674 (0.95)	13,354 (1.00)
Total	3,202 (0.08)	36,577 (0.92)	39,779 (1.00)	11,105 (0.04)	267,887 (0.96)	278,992 (1.00)

Note: 1. The weights are the number of "likes" and "dislikes" that other readers expressed on the comments.
2. The numbers in parentheses indicate the proportion.

4. 댓글과 설문의 주제 비교 분석

토픽모형(topic modeling)은 텍스트의 주제를 요약하는 기법으로 텍스트의 차원을 축소하는 비확률적 과정(주성분분해, 특이값분해, 잠재의미분석)과 확률적 과정인 LDA(Latent Dirichlet Allocation) 등이 있다(김수현 외, 2020). 특히 LDA는 베이지안 확률과정으로 임의의 주제에 특정 키워드가 나타날 조건부 확률로 주제를 추정

29) 김은미·이준웅(2006)이 주장하는 소수 활동가의 지배 현상과 같은 맥락이다.

하는 토픽모형이다. 여기에서는 LDA로 댓글과 설문에 담긴 각각의 키워드를 중심으로 주요 주제를 추정하고 그 차이점과 공통점을 비교분석한다.

LDA(Latent Dirichlet Allocation)은 문서에 사용된 단어를 기반으로 예상되는 주제에 대한 확률을 계산하는 과정이다. 문서집합 전체 문서 개수를 D, 사전 정의된 토픽의 수 K, d번째 문서의 총 단어 개수 N, d번째 문서의 n번째 단어 $w_{d,n}$, 디리클레 분포의 모수 α, β가 주어졌을 때, ϕ_k를 각 단어들이 k번째 토픽에 속할 확률벡터라고 하고, θ_d를 d번째 문서가 가진 토픽 비중을 나타내는 벡터라고 하면 LDA에서는 전체 문서 내 한 단어가 쓰일 조건부확률 $P(w_{1:D} \mid \phi_{1:K}, \theta_{1:D}, z_{1:D}, ;\alpha, \beta)$이 ① 각 토픽의 말뭉치 내 단어에 대한 분포 $\prod_{i=1}^{K} P(\phi_i \mid \beta)$와 ② 각 문서가 지닌 토픽분포 $\prod_{i=1}^{D} P(\theta_d \mid \alpha)$의 결합분포로 나타난다. 따라서 LDA는 문서의 한 단어가 관찰되었을 경우 ①과 ②를 아래와 같이 베이지안 확률과정으로 추정[30]하는 과정이다.

$$P(\phi_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{P(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{P(w_{1:D})}$$

$$P(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}; \alpha, \beta)$$
$$= \prod_{i=1}^{K} P(\phi_i \mid \beta) \prod_{i=1}^{D} P(\theta_d \mid \alpha) \left\{ \prod_{i=1}^{N} P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid \phi_{1:K}, z_{d,n}) \right\}$$

한편 토픽의 개수는 임의로 정하기보다 아래와 같이 통일성 점수(coherence score)[31]를 통해 가장 높은 점수를 부여받은 토픽 개수를 기준으로 한다. 여기에서 w_i와 w_j는 문서내 단어집합(W)에 속한 단어이며 통일성 점수(C)가 높을수록 바람직한 토픽의 개수라고 평가할 수 있다.

$$C(W) = \left\{ \sum_{w_i \in W} \left(\frac{\log \frac{P(w_i,\ w_j) + \epsilon}{P(w_i)\,P(w_j)}}{-\log(P(w_i,\ w_j) + \epsilon)} \right)^{\gamma} \right\}_{j=1,\ldots,N}$$

30) Gibbs 샘플링과 Bayesian 추정을 반복 시행.

31) 자세한 내용은 Roder et al. (2015) 참조.

1) 기사와 댓글의 주제

〈Table 6〉은 LDA 추정을 통해 연도별 기사에서 추출한 주제이며 추정결과는 키워드로 구성되어있으나 키워드를 읽는 것만으로도 어떠한 주제인지 짐작할 수 있을 정도로 연관성 있는 단어로 구성되어 있다.[32] 2018년의 경우 기사의 수도 적었지만 주제도 단지 가상자산에 대한 과세에 국한되지 않고 부동산, 주식 등 보유자산 수익에 대한 포괄적 과세에 대한 다양한 논의의 일환이었음을 알 수 있다. 2019년에는 가상화폐거래소인 '빗썸'과 같은 거래소에 과세를 하는 방안이 국회 등에서 논의되고 있다. 또한 암호화폐 거래에 대한 과세를 거래소자료를 통해 원천징수하는 방안 역시 논의되었다는 점을 짐작할 수 있다. 2020년 이후 주제도 다양해지는 동시에 과세에 대한 논의도 구체적으로 진행되었다. 이 시기에는 가상자산 과세근거, 비거주자의 가상자산 국내 거래 시 세법 적용 문제, 세원을 정확히 파악할 수 있는 기술의 부재 속에 신고방식 도입 가능성 등에 대한 논의와 협의가 진행되고 있었다는 점도 확인 가능하다. 2021년에도 2020년과 유사한 논의가 이어지고 있었으나 대선을 앞두고 일시적으로 과세를 유예하였다는 점을 알 수 있다. 2022년에는 가상자산 과세에 대한 좀 더 다양한 논의가 지속되고 선거기간 중 일시 유예되었던 과세방침이 새 정부 출범과 동시에 재검토에 들어간 사실도 주제에 드러나 있다.

한편 주제의 분류에서 볼 수 있듯이 뉴스 기사들의 경우에는 가상자산 기타소득세에 대한 납세자들의 인식보다는 국회나 기획재정부에서 조세정책관련 법률을 검토하는 사실을 전달하거나, 가상자산의 기술적 근간이 되는 블록체인 기술을 소개하거나, 대권주자들의 공약을 전달하는 주제가 많은 것을 확인 할 수 있다. 예를 들어 2018년의 경우 28%의 주제가 블록체인 기술에 관련된 것이었으며 2022년의 경우 44%가 대선 후보와 관련이 있는 주제였다.

〈Table 7〉은 LDA 추정을 통해 연도별 기사의 댓글에서 추출한 주제이다. 추정결과는 키워드로만 구성되어 있지만 대체적으로 조세정책 관련 정부정책에 대한 과격한 불만을 토로하는 내용인 것을 알 수 있다. 중요한 것은 가상자산 과세 관련 기사의 댓글에서 부동산 과세 관련 주제도 상당히 많이 추출된다는 것이다. 예를 들어 2018년의 경우 가상자산 과세 관련 댓글에서도 실질적으로 부동산 과세에 대

32) 본문의 표에서는 주제별 단어의 수를 10개 이하로 제한하였다.

〈Table 6〉 Topic Modeling Results of News Articles on Virtual Asset Taxation

Year	Keywords (Topics)	Article Counts	Proportion
2018	income, currency, stocks, profits, investors, levy, income tax	13	0.41
	deputy prime minister, planning, issues, National Assembly, information, policy, Ministry of Finance, discussion	10	0.31
	blockchain, technology, market, business, bitcoin, real estate, holdings, funds	9	0.28
		(32)	(1.00)
2019	Bithumb, foreigner, source, collection, industry, customer, profit	18	0.33
	currency, price, market, South Korea, holdings	17	0.31
	National Assembly, next year, policy, passage, Ministry of Finance, revision, news, information	19	0.35
		(54)	(1.00)
2020	transaction tax, blockchain, tax, arbitrage, policy, method, tax, opinion, association	24	0.20
	Bithumb, National Tax Service, basis, income tax law, collection, source, object, non-resident	20	0.17
	investment, investor, market, digital, stock, service, announcement	23	0.19
	amount, classification, acquisition, description, inclusion, value, review, nature, margin, policy	27	0.23
	business, reporting, industry, enforcement, law, funding, establishment, verification, obligation, account	26	0.22
		(120)	(1.00)
2021	news, economy, protection, chairman, institution, Korea, recognition, law	114	0.21
	equity, deduction, other, standard, application, margin, income tax, amendment, case	99	0.18
	bitcoin, coin, price, report, information, operator, dollar, National Tax Service	154	0.28
	Congressman, Democratic Party, probation, deputy prime minister, policy, position, people, candidate, discussion	178	0.33
		(545)	(1.00)
2022	investigation, presidential election, representative, public opinion, Jaemyung Lee, real estate, news, generation, postponement, election	11	0.15
	income, tax deduction, income tax, parliament, box, apply, scheduled, law, enforcement	14	0.19
	bitcoin, dollar, price, quote, US, year, outlook, gift, evaluation	14	0.19
	blockchain, business, metaverse, problem, platform, technology, work	14	0.19
	government, protection, public, Yoon Seok-yeol, disclosure, limit, review, permitting, preparation	22	0.29
		(75)	(1.00)

Note: Keywords are limited to 10 or less, and the number of topics is determined based on the coherence score. The numbers in parentheses represent the sum of counts and proportions.

〈Table 7〉 Topic Modeling Results of Comments on Virtual Asset Taxation

Year	Keywords (Topics)	Comment Counts	Proportion
2018	country, investment, words, people, real estate, money	867	0.44
	currency, money, virtual, coin, person, taxation, stock, exchange, regulation, us	572	0.29
	tax, government, thought, recognition, common people, exchange, closure	552	0.28
		(1,991)	(1.00)
2019	stock, country, yours, regime, thoughts, investment, economy, market,	801	0.35
	tax, money, government, person, people, door, disaster, levy, dog	784	0.34
	tax, currency, recognition, bitcoin, taxation, virtual, income, coin	697	0.31
		(2,282)	(1.00)
2020	stock, country, market, ant, taxation, transfer tax, finance, transaction tax, us	776	0.22
	tax, horse, democratic party, one, korea, dog pig	752	0.21
	thoughts, real estate, policy, investment, bitcoin, economy, country	602	0.17
	currency, virtual, exchange, coin, recognition, asset, crypto	751	0.21
	government, people, regime, person, tax, dog, income	624	0.18
		(3,505)	(1.00)
2021	country, people, Democratic Party, regime, election, time, vote, policy, next year, presidential election	8,660	0.29
	government, thoughts, words	5,876	0.20
	person, bitcoin, coin, real estate, speculation, market, gambling	4,938	0.17
	tax, currency, virtual, protection, government, asset, recognition, country	5,355	0.18
	coins, stocks, taxation, investment, exchanges, income	4,562	0.16
		(29,391)	(1.00)
2022	virtual, bitcoin, currency, taxation	926	0.43
	tax, stock, pledge, abolition, investment, transfer tax, short sale, dog	645	0.30
	coin, thing, person, money, country, people, Jaemyung Lee, president, Democratic Party, Seokyeol Yoon	589	0.27
		(2,160)	(1.00)

Note: Keywords are limited to 10 or less, and the number of topics is determined based on the coherence score. The numbers in parentheses represent the sum of counts and proportions.

한 주제를 지닌 댓글의 비율(44%)로부터 과세관련 반대의견은 사실 종합부동산세에 대한 과세에 대한 불만을 표출한 것임을 알 수 있다. 2019년 댓글도 34%는 전반적인 조세부담 증대에 대한 불만을 주제로 하고 있다. 2022년의 경우도 30%의 댓글은 금융투자소득세와 공매도에 관한 내용을 주제로 하고 있으며 27%의 경우에는 대선 후보들에 대한 내용을 주제로 하고 있다. 따라서 인터넷 미디어 여론인 댓글은 가상자산 과세에 대한 인식뿐만 아니라 조세 정책 전반에 대한 의견이나 정권에 대한 평가 등 실질적으로는 주요 논점에서 벗어난 주제들이 많이 섞여있다는 것을 알 수 있다.

2) 설문 답변의 주제

〈Table 8〉은 3회에 걸쳐서 이루어진 가상자산 기타소득세 관련 개방형 설문에서 긍정과 부정의 응답의 주제를 LDA로 분석한 결과이다. 응답 중 적절하지 않은 답변을 제외한 수정표본에서 주제를 추출하였으며 긍정과 부정 응답에서 각각 7개의 주제가 추출되었다. 긍정 답변의 경우 주제별로 응답을 분류하였을 경우 해당 토픽에 해당하는 응답이 12~20% 비율로 고르게 분포된 것으로 나타나며, 부정의 경우는 각 토픽이 8~20% 비중을 차지하는 것으로 나타났다.

긍정답변의 경우 20%에 해당하는 비율로 나타난 '화폐, 돈, 수익, 사람, 가치, 이득, 자금' 키워드를 지닌 주제는 수익에 대해 당연히 과세를 해야 한다는 논리로 해석 할 수 있다. 두 번째 주제는 '무분별, 투자, 규제, 투기, 범죄, 방지, 조세, 제도' 키워드로 구성되어 있으며 무분별한 투자를 규제하고 투기와 범죄를 방지하는 목적의 조세제도로 가상자산 과세를 긍정적으로 생각하는 주제임을 알 수 있다. 세 번째 주제는 '불로소득, 일반, 개인, 소득세, 세율, 적용, 원칙, 동일'로 가상자산 투자에 의한 소득이 불로소득이며 과세율 20%는 소득세율에 비추어볼 때 과하지 않다는 입장인 것을 알 수 있다. 네 번째 주제는 '주식, 재산, 경우, 세금, 부과, 해당, 이유, 부여, 금액'의 키워드로 구성되어 주식 거래의 양도세에 비추어볼 때 가상자산 과세는 형평성에 맞는다는 주장인 것으로 볼 수 있다. 다섯 번째부터 일곱 번째 주제까지는 키워드만으로 뚜렷한 주제를 특정 짓기 어렵다. 그러나 가상자산 과세가 필요하며 공평하다고 생각하는 주제임을 알 수 있다.

부정응답의 주제도 긍정응답과 일부 키워드를 공유하고 있으나 다른 논리에 의해

〈Table 8〉 Topic analysis of positive/negative responses from the open survey on virtual asset taxation

	Keywords (Topics)	Counts	Proportion
Positive Responses	currency, money, revenue, people, value, gain, funds	271	0.20
	indiscrimination, investment, regulation, speculation, crime, prevention, taxation, system, equity	209	0.15
	unearned income, general, individual, income tax, tax rate, application, principle, same	195	0.14
	stock, property, case, tax, levy, applicable, reason, grant, amount	186	0.14
	thoughts, positive, government, people, fairness, cash, portion, work	171	0.13
	taxation, income, need, economy, possible, kind, policy	168	0.12
	transaction, profit, accrual, tax evasion, due, country, taxation, use, recognition	164	0.12
		(1,364)	(1.00)
Negative Response	risk, investment, money, means, taxation, injustice, idea, value, due, extent	126	0.20
	virtual, asset, real, market, real money, variable, income tax	83	0.13
	virtual, money, law, recognition, damage, stability	84	0.14
	investor, loss, protection, institution, taxation, government, regulation, policy	73	0.12
	tax, levy, transition, country, revenue, portion	120	0.19
	tax rate, due, fraud, application, individual, burden, profit, occurrence	82	0.13
	income, assets, virtual, stocks, trading, finance, case, basis, equity	49	0.08
		(617)	(1.00)

Note: Keywords are limited to 10 or less, and the number of topics is determined based on the coherence score. The numbers in parentheses represent the sum of counts and proportions.

반대 의견이 서술되고 있다. 첫 번째 주제는 '위험, 투자, 돈, 수단, 과세, 부당, 생각, 가치, 때문, 정도' 키워드로 구성되어있으며 위험을 감수한 투자에 과세하는 것은 부당하다는 내용임을 알 수 있다 두 번째 주제 키워드는 '가상, 자산, 실물, 시장, 실제 현금, 변동, 소득세'로 현금과 같은 실물 자산과 달리 변동성이 큰 자산으로 가상자산에 과세하는 것이 부당하다는 주제의 응답임을 알 수 있다. 세 번째

주제는 '가상, 화폐, 법, 인정, 손해, 안정'으로 가상자산이 제도적으로 자산으로 인정되어야 과세가 정당하다는 의견이다. 네 번째 주제는 '투자자, 손실, 보호, 제도, 과세, 정부, 규제, 정책'로 시스템 리스크에 의한 투자손실을 보호하는 제도적 장치 미비에도 불구하고 정부가 지나치게 규제 위주의 정책을 펼친다는 주제의 응답일 가능성이 높다. 다섯 번째 주제는 '세금, 부과, 과도, 국가, 수익, 부분'으로 20%의 세율이 과하다는 응답일 수 있다.

이상에서 확인할 수 있듯이 가상자산 기타소득세 관련 개방형 설문에서는 해당 정책에 밀접하게 연관된 주제들로 구성되어 있다. 이를 바탕으로 볼 때 개방형 질문에 대해서 응답한 텍스트는 조세정책에 대한 선호를 잘 반영하고 있을 것으로 판단된다. 반면 앞에서 분석한 댓글의 경우 주제가 가상자산 과세뿐만 아니라 종합부동산세, 정당 등에 대한 반대의견이 혼재되어있다. 따라서 댓글의 경우 주제의 일관성 측면에서도 여론 대표성이 있는지 여부를 의심해볼 만하다.

위와 같은 분석에도 불구하고 본 연구의 토픽모형에는 한계가 존재한다. LDA는 비모수적 추정 방법이므로 비모수적 접근방법의 내재적 한계로부터 LDA 역시 이로부터 자유로울 수는 없다. 우선 LDA의 분석결과는 키워드 집합과 각 키워드의 조건부 확률로 제시된다. 따라서 이를 하나의 주제로 분석하는 것은 연구자의 몫이다. 키워드 집합이 일관되게 하나의 주제를 나타내고 있는 경우 임의의 주제로 분석하는 것에 문제는 없으나 키워드가 일관되지 못한 경우에는 연구자의 해석이 개입될 수밖에 없다. 그러므로 본 연구에서도 분석 결과 제시된 키워드에 따라 어떠한 논의가 있었을 것인지를 추측하였을 뿐, 연구자가 직접적으로 어떠한 주제임을 명시하지는 않았다. 한편 키워드 집합이 일관성 있게 제시되지 않은 경우 키워드만으로 주제를 특정하기 어려움을 나타내었다.

V. 결 론

인터넷 여론은 우리가 가장 쉽고 보편적으로 접할 수 있는 여론의 장이다. 인터넷 여론의 접근성과 개방성을 바탕으로 정책당국 또는 국회와 국민 간 활발한 의사소통이 이루어질 수 있는 수단이 마련되었다는 점에서 인터넷 여론의 긍정적인 역할을 기대할 수 있다. 그럼에도 불구하고 인터넷 여론이 전체 여론을 대표할 수 있는 매체라는 인식은 사실과 다를 수 있다.

본 연구에서는 인터넷 미디어 여론이 조세정책에 대한 국민적 선호를 잘 반영하고 있는지를 텍스트 마이닝 방법론을 통해 정성적, 정량적으로 분석해 보았다. 이를 위해서 먼저 개방형 설문을 통해 가상자산 기타소득세에 대한 긍정/부정 여부와 그렇게 생각하는 이유에 대해 설문하여 실제 국민들의 선호와 근접한 텍스트를 수집하였다. 그리고 이렇게 얻어진 텍스트를 인터넷 미디어 여론인 댓글과 긍정/부정 비율 및 주제 측면에서 정량적으로 비교해 보았다. 그 결과를 요약하자면 특정 조세정책에 대한 인터넷 미디어 여론은 텍스트의 주제가 매우 산발적이어서 조세정책에 대한 선호를 잘 반영하지 못할 것으로 판단할 수 있다. 또한 감성분석 결과는 인터넷 미디어 여론이 조세정책에 대한 부정적인 의견을 매우 편향되게 표출하여 조세정책에 대한 국민들의 선호를 왜곡하여 반영할 가능성이 있음을 정량적으로 보여주고 있다. 이는 인터넷 여론을 의식하여 정책을 수정하거나 철회하는 것은 사회적 합의를 번복하는 것으로 공정하지 못하며 세제안정성을 저해할 수도 있다는 주장을 뒷받침하는 결과이다. 이는 정책결정자들이 인터넷 여론을 의식하여 성급히 정책을 결정하거나 수정하기보다는 조세정책에 대한 국민적 선호를 충분히 논의할 수 있는 논의의 장을 마련하고 국민적인 합의를 통해서 조세정책을 운용할 필요가 있음을 시사한다.

본 연구가 조세정책에 대한 인터넷 댓글의 여론 대표성을 검증하는 최초의 연구라는 점에서 기여하는 바가 있음에도 불구하고 분명한 한계와 개선방향에 대해 언급하고자 한다. 우선 앞서 언급한 토픽모형의 한계이며 이를 극복하기 위해서는 최근 발표된 거대 인공지능 언어모형을 활용한 토픽모형으로 분석해야 한다. 거대 인공지능 언어모형의 경우 본 연구에서 활용한 베이지안 기법뿐만 아니라 강화학습(reinforcement learning)을 통해 키워드의 일관성을 높일 수 있다. 따라서 분석결과로 키워드와 조건부 확률이 제시되는 수준을 넘어 요약된 주제를 제시할 수 있다. 이러한 거대 인공지능 언어모형을 활용한 토픽분석은 차후의 연구주제로 남겨둔다. 또한 가상자산 과세에 대한 인터넷 댓글과 설문조사를 분석하였다는 점에서, 본 연구의 분석 결과를 다른 조세 관련한 정책 분석에 단순 확대 적용하는 데는 주의할 필요가 있다. 본 연구는 가상자산 과세라는 특정 조세정책에 대한 대중의 반응을 분석한 연구이므로 일반적인 조세정책에 대한 대중의 반응을 분석한 결과로 의제할 수는 없을 것이다. 일반적인 조세정책에 대한 인터넷 여론대표성에 대한 연구도 차후 연구과제로 남겨둔다.

■ 참 고 문 헌

1. 국회예산정책처, 『2013 세법개정안 분석』, 국회예산정책처, 2013.

   (Translated in English) National Assembly Budget Office, *Analysis of the 2013 Tax Law Amendment*, National Assembly Budget Office, 2013.

2. _____, 『트렌드 세법 -과거 20년간 세법개정안의 궤적을 담다』, 국회예산정책처, 2017.

   (Translated in English) National Assembly Budget Office, *Trend Tax Law: The Trajectory of Tax Law Amendment Bills over the Past 20 Years*, National Assembly Budget Office, 2017.

3. 경향신문, "대선 앞두고…세금 깎아줄게, 표 좀 다오?" 2021.

   (Translated in English) The Kyunghyang Shinmun, "Before the Presidential Election... Will You Vote for Me if I Cut Your Taxes?" 2021.

4. 기획재정부, "5,500만원 이하자 세법개정에 따른 세부담 증가 해소 -2015년 연말정산 보완대책-," 기획재정부 보도자료, 2015.

   (Translated in English) Ministry of Economy and Finance, "Relief of Increased Tax Burden for Those with Annual Income below KRW 55 Million due to Tax Reform - Supplementary Measures for Year-End Tax Settlement in 2015," Press Release by the Ministry of Economy and Finance, 2015.

5. 김수현 · 이영준 · 신진영 · 박기영, "거시경제 분석을 위한 텍스트 마이닝," 『한국경제의 분석』, 제26-1집, 2020, pp. 1-70.

   (Translated in English) Kim, S., Y. Lee, J. Shin, and K. Park, "Text Mining for Macroeconomic Analysis," *Journal of Korean Economic Analysis*, Vol. 26, No. 1, 2020, pp. 1-70.

6. 김은미 · 이준웅, "읽기의 재발견: 인터넷 토론 공간에서 커뮤니케이션의 효과," 『한국언론학보』, 제50-4집, 2006, pp. 65-94.

   (Translated in English) Kim, E., and J. Lee, "Rethinking 'Reading'online : The Effects of Online Communication," *Korean Journal of Journalism and Communication Studies*, Vol. 50, No. 4, 2006, pp. 65-94.

7. 박지웅 · 김재진 · 구재이, 『세금, 알아야 바꾼다』, 메디치미디어, 2018.

   (Translated in English) Park, J., J. Kim, and J. Koo, *Taxes, We Need to Know to Change*, Medici Media, 2018.

8. 배진수 · 박정흠 · 김수현, "텍스트 분석을 이용한 조세정책에 대한 인식 연구," 한국조세재정연구원, 2022.

   (Translated in English) Bae, J., C. Park, and S. Kim, "A Text-analysis on the Perception of Taxation: The Case of Virtual Asset tax in Korea," Korea Institute of Public Finance, 2022.

9. 임원혁 · 이창근 · 최동욱 · 정세은, "한국의 여론양극화 양상과 기제에 관한 연구," 한국개발연구원, 2019.

   (Translated in English) Lim, W., C. Lee, D. Choi, and S. Jung, "Opinion Polarization

in Korea: Its Characteristics and Drivers," Korea Development Institute, 2019.

10. 최동욱, "인터넷 포털의 경쟁과 뉴스 콘텐츠의 선택," 정책연구시리즈 2017-01, 한국개발연구원, 2017.

    (Translated in English) Choi, D., "Internet Portal Competition and Economic Incentive to Tailor News Slant," Policy Research Series 2017-01, Korea Development Institute, 2017.

11. 한겨레, "국민 2명 중 1명은 네이버로 뉴스 본다," 2022, https://www.hani.co.kr/arti/economy/it/1067738.html (접속일자 2023.04.20.).

    (Translated in English) The Hankyoreh, "One out of Two Koreans Reads News on Naver," 2022.

12. 황남석, "조세법률주의의 역사적 계보,"『사법』, 제1집 제38권, 2016, pp.123-168.

    (Translated in English) Hwang, N., "Historical Genealogy of the Principle of No Taxation without Law," *JURIS*, Vol. 1, No. 38, 2016, pp.123-168.

13. 황의찬 · 우석진, "납세자의 정서가 정부의 조세정책 결정에 미치는 영향,"『법경제학연구』, 제19집 제1권, 2022, pp.229-263.

    (Translated in English) Hwang, E., and S. Woo, "The Effect of Taxpayer Sentiment on Government Tax Policy Decisions," *Korean Journal of Law and Economics*, Vol. 19, No. 1, 2022, pp.229-263.

14. Baker, S. R., N. Bloom, and S. J. Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, Vol. 131, No. 4, 2016, pp.1593-1636.

15. Buckman M., and A. Joseph, "An Interpretable Machine Learning Workflow with an Application to Economic Forecasting," Bank of England Staff Working Paper No. 984, 2022.

16. Gentzkow, M., B. Kelly, and M. Taddy, "Text as Data," *Journal of Economic Literature*, Vol. 57, No. 3, 2019, pp.535-574.

17. Gentzkow, M., and J. M. Shapiro, "What Drives Media Slant? Evidence from US Daily Newspapers," *Econometrica*, Vol. 78, No. 1, 2010, pp.35-71.

18. Groseclose, T., and J. Milyo, "A Measure of Media Bias," *Quarterly Journal of Economics*, Vol. 120, No. 4, 2005, pp.1191-1237.

19. Gould, A. C., and P. J. Baker, "Democracy and Taxation," *Annual Review of Political Science*, Vol. 5, No. 1, 2002, pp.87-110.

20. Daude, C., H. Gutiérrez, and Á. Melguizo, "What Drives Tax Morale?" OECD Development Center Working Paper No. 315, 2012.

21. Hansen, S., and M. McMahon, "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communications," *Journal of International Economics*, Vol. 99, No. 1, 2016, pp.114-133.

22. Jun, B. H., M. Cho, and M. H. Park, "Procedural Fairness and Taxpayers' Response: Evidence from an Experiment," *Korean Economic Review*, Vol. 31, No. 2, 2015, pp.301-326.

23. Lee, Y., ekorpkit: eKonomic Research Python Toolkit, https://zenodo.org/record/6592044#.ZDnSGHZBxD-/, 2022.

24. Lee, Y., S. Kim, and K. Y. Park, "Deciphering Monetary Policy Committee Minutes with Text Mining Approach: The Case of South Korea," *Korean Economic Review*, Vol. 35, No. 2, 2019, pp. 471-511.

25. _____, "Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea," Bank of Korea WP 2019-11, 2019.

26. Lucca, D., and F. Trebbi, "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements," NBER Working Papers No. 15367, 2011.

27. Nyman, R., S. Kapadia, and D. Tuckett, "News and Narratives in Financial Systems: Explotting Big Data for System Risk Assessment," *Journal of Economic Dynamic and Control*, Vol. 127, 2021.

28. Picault, M., and T. Renault, "Words Are Not All Created Equal: A New Measure of ECB Communication," *Journal of International Money and Finance*, No. 79, 2017, pp. 136-156.

29. Roder, M., A. Both, and A. Hinneburg. "Exploring the Space of Topic Coherence Measures," WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 399-408.

30. Torgler, B., *Tax Compliance and Tax Morale: A Theoretical and Empirical Analysis*, Edward Elgar Publishing, 2007.

31. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L, Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," arXiv:1706. 03762, 2017.

32. Zhao, W. X., J. Jiang, J. Weng, J. He, E. Lim, P. H. Yan, and X. Li, "Comparing Twitter and Traditional Media using Topic Models," *European Conference on Information Retrieval*, 2011, pp. 338-349.

〈 Appendix 〉

〈Table A-1〉 Demographics in Samples of the Tax Perception Survey

(Unit: %)

Demographics		1st Survey	2nd Survey	3rd Survey
Age	10s	6.7	4.2	5.1
	20s	17.3	18.2	17.4
	30s	19.6	19.7	19.8
	40s	19.6	19.3	19.8
	50s	20.3	21.7	21.0
	60s and above	16.5	16.8	16.9
Sex	Male	51.0	51.8	52.1
	Female	49.0	48.2	47.9
Region	Seoul	32.7	33.1	32.2
	Gyeonggi	22.0	22.4	23.0
	Incheon	6.4	5.9	5.8
	Daejeon	2.8	2.8	2.4
	Sejong	1.0	0.4	0.6
	Chungnam	1.8	2.1	1.6
	Chungbuk	1.1	1.1	1.1
	Gwangju	2.7	2.8	3.2
	Jeonnam	0.7	1.0	1.1
	Jeonbuk	2.9	3.0	3.2
	Daegu	5.3	5.2	5.5
	Gyeongbuk	3.2	3.4	3.1
	Busan	7.3	7.5	7.5
	Ulsan	1.7	1.7	1.4
	Gyeongnam	6.2	5.4	5.5
	Gangwon	1.5	1.3	2.0
	Jeju	0.7	0.8	0.7
Occupation	Professionals	8.8	9.2	10.7
	Self-employed	10.1	11.0	10.6
	Students	10.2	8.9	7.8
	Homemakers	11.0	9.9	9.6
	Unemployed	7.0	8.2	7.3
	Employees	52.9	52.9	54.0
Monthly Income	Less than 1 million won	3.9	3.7	3.2
	100-199 million won	5.0	4.2	4.1
	200-299 million won	14.7	11.1	13.6
	300-399 million won	13.3	16.2	17.1
	400-499 million won	16.6	15.2	14.3
	500-599 million won	13.7	17.5	14.5
	600-699 million won	11.5	10.6	9.7
	700 million won or more	21.3	21.4	23.4
Education	Less than middle school	2.7	1.7	1.4
	High school graduate	22.2	20.0	20.2
	College graduate	66.9	67.3	69.1
	Post-graduate	8.3	11.0	9.3

〈Table A-2〉 Responses Excluded among 'Positive'

	Reponses Excluded	
Positive	없음,	너무좋아서여
	신선함이 가득함,	가격이저렴해서좋다
	가상자산,	돈을 많이 번다
	모르겠음,	그냥 저냥
	굿굿 ,	가상자산을 안해서
	창의적인 서비스입니다,	별다른 이유는 없다.
	정놀	잘 알지 못해서
	다양하고 트렌디하다	좋아서
	현재는 그에대한불만이없음	기술력 진보
	이용이 쉽다	투자
	좋으니까	그냥
	스토리	자산증가
	손쉽게양도가능	과세열거방식
	펼리할거 같다	과세방안
	돈의가치라생각	없습니다
	좋다고 생각한다	잘 모르겠다
	특별한 의견이 없습니다	투자해보았다
	편하다	발전 가능성이 있다
	글세요	다양한 소득
	잘모름	어어어어어ㅓ아어어으
	너무나도 없는 것 같다	왠지 바쁜일상에 편리할듯합니다
	명분	원래있었던데로
	잘모르겠음	간편하다
	모름	안정적이라서
	새로운 화폐	그냥
	신세대적이다	관심 없음스토리
	편리한점이 기억에 남음	그냥 . . ?
	취득세	특별한 생각없음
	난하지않으니까	도있는사람들이하니까
	흥미롭다	매우 경제적이어서
	그냥그랬습니다	글세요. . . . .
	긍정적인 생각	긍정
	좋은거 같아서	잘 모르겠음
	똑같은 돈이다.	모르겠음
	긍정 신세계	비싸다
	글세요	효율성이 있다
	소신적인 자발적 참여	너무 좋아요
	좋아요	처음들어봄
	코인이 없는 것 같다.	너무좋아서여
	특별히 없음	좋습니다
	이유 없음	글세요
	그만큼 유용해졌으니따	신 화폐
	편하고 이대로 쭉	나쁘게생각해본적이없다
	가격이저렴해서좋다	특별한 의견이 없습니다.
	없습니다	나는 안해서잘모르겟즘
	좋습니다	재산 증식을위한
	딱히 없다.	세금 부여

〈Table A-3〉 Responses Excluded among 'Negative'

	Reponses Excluded	
Negative	그냥 특별히 없음	투기 자산
	잘모르겠음	굳이?
	굳이인느낌	잘모름
	신기루	가상
	잘 모른다	특별히 없음
	..... 좋습니다	탈세
	모르겠음	투기 조작
	투기	애매하다
	세금	잘몰라요.....
	재산이다	투기성
	더 올려야 한다	그냥
	특별히 없음	모르겠음
	좋습니다	

〈Table A-4〉 Negative Comments Misclassified as Positive

	Comments	Classification probability		Results
		P	N	
1	통화로 인정하진 않지만 세금은 걷겠다는 뜻.	0.87	0.23	P
2	다 나락인디 수익이 나야 세금을 내지	0.90	0.10	P
3	조만간 네이버 지식인 내공도 싸이월드 도토리도 보유세 낼듯	0.91	0.09	P
4	비과세가 딱이야!	0.57	0.43	P

Note: Four examples of negative comments misclassified as positive.

〈Table A-5〉 Cases correctly Classified as Positive or Negative

	Comments	Classification probability		Results
		P	N	
Positive				
1	과세은 해야지..	0.93	0.07	P
2	이건 잘했네. 미래는 디지털 경제다. 미래 먹거리.	0.93	0.07	P
3	요게 맞지. 주식이야 기업에 투자 개념이지만 코인은 사실상 투기인데 이런 자산에 비과세 혜택을 주는건 이상한거지	0.91	0.09	P
4	만약 코인과세를 하지 않으면... 코인은 제도권으로 들어오지 않고, 피해자들이 생긴다. 그리고, 세금을 월급쟁이들이 다낸다	0.96	0.04	P
5	소득있는곳에 세금있는건 당연한거지	0.97	0.03	P
Negative				
1	가상화폐 손실나면 얼마나 보상해줄건데?	0.04	0.96	N
2	가상자산 과세 1년에 10억원 이상 벌었을 때만 10% 과세 해라	0.14	0.86	N
3	예금자보호하고 루나같은사태때 어찌할것인지 먼저구축해놓고 과세해라	0.03	0.97	N
4	세금 받을꺼면 코인투자자들도 손실보장해줘라 ㅋ	0.05	0.95	N
5	쪼대로 하고 손해 보면 보전 해줄거냐고.	0.03	0.97	N

Note: Examples of negative comments classified as negative and positive comments classified as positive.

Evaluation of Representation of Policy Public Opinion in Internet Comments: A Case of Cryptocurrency Taxation*

Soohyon Kim** · Jinsoo Bae***

Abstract

We evaluate whether the contents of comments on internet articles represent public opinion on tax policy through the case of income tax on cryptocurrency investment. Internet articles related to tax policy are easily accessible to anyone, and opinions can be freely expressed in comments. Therefore, it can be regarded that the comments on the article represent public opinion on the policy. In this study, based on text data collected by surveys through random sampling, we train a transformer model to build a sentiment analyzer that can classify positive/negative opinions on the new income tax policy on cryptocurrency and obtain the positive/negative ratio for the comments. As a result, we find that the rate of negative opinions in the policy responses to the comments on the articles was 80 to 90% of the total comments. It can be seen that this shows a considerable level of bias compared to the negative rate of about 30% of the open-ended survey through random sampling. The results shed light on the process of collecting public responses to government policies such as tax policies for determining whether or not to accept comments on articles as evidence of public opinion.

Key Words: comments on internet article, internet public opinion, tax policy, sentiment analysis, deep learning
JEL Classification: C8, H2, H3

Received: Feb. 15, 2023. Revised: May 4, 2023. Accepted: May 23, 2023.

* I would like to extend my gratitude to the two anonymous reviewers who provided meticulous critiques to greatly enhance the structure and quality of this paper. I would also like to express my appreciation to Dong-Ik Kang, Changsu Ko, and Jeonghwan Kim at the Korea Institute of Public Finance, who gave valuable and useful advice for this study, and to the participants in the special session of the 2023 Korea's Allied Economic Associations Annual Meeting. My deepest thanks go to Professor Youngjoon Lee from Jeju Halla University who provided technical advice related to the model. This study is a revised and supplemented version of a portion of the basic research project 'A Study on Tax Compliance Using Text Analysis' conducted by the Korea Institute of Public Finance in 2022. This paper was presented at the 2023 Korean Journal of Economic Studies Special Session (applied economic research using big data, unstructured data, and AI) at the 2023 Korea's Allied Economic Associations Annual Meeting. Any errors remaining in the paper are the responsibility of the authors.

** First Author, Assistant Professor, Department of Economics, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Korea, Phone: +82-62-530-1548, e-mail: soohyon.kim@jnu.ac.kr

*** Corresponding Author, Associate Fellow, Fiscal System Analysis Team, Korea Institute of Public Finance, 336 Sicheong-daero, Sejong 30147, Korea, Phone: +82-44-414-2440, e-mail: jsbae@kipf.re.kr