제36회 기독교학문학회 발표논문 (19.10.26)

교수일반분과/경영B(로고스)/02/

토픽 모델링을 이용한 유튜브 설교 동영상 리뷰 분석*

임병학**

<요 약>

다양한 대용량의 데이터를 처리하는 컴퓨터 기술의 발달로 텍스트 데이터를 분석하는 방법은 기계학습과 자연어처리에서 최근 토픽모데링으로 큰 발전을 이루어 왔다. 토픽모델링은 문서에서 추상적인 "주제"를 발견하기 위한 모델로 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용하는 텍스트 마이닝 기법 중 하나이다.

본 연구는 유튜브 설교 동영상의 리뷰를 토픽 모델링으로 설교에 대한 반응을 몇 개의 토픽별로 나누어 분석하였다. 분석결과 5개의 토픽으로 분류하였다. 토픽 5개 중 3개 토픽 1, 2, 4는 온라인 고객 리뷰들처럼 본 설교 동영상의 핵심 속성으로 볼 수 있지만, 다른 두 토픽 3과 5는 설교 동영상과는 거리가 있는 리뷰가 포함된 것을 볼 수 있다. 이는 설교 동영상보다는 설교자의 이전 말이나 행위들에 대해 이야기한 리뷰들이라고 볼 수 있다. 따라서 교회 입장에 볼 때 리뷰에 대한 체계적이고 객관적인 평가와 관리가 요구된다. 사용자들이 기록한 리뷰들을 토필별로 분류하여 긍정적인 리뷰들과 가장 관련성이 높은 리뷰들을 부각시키는 리뷰의 재배열이 필요하다. 또한 긍적적인 리뷰들을 포함한 토픽을 사용하여 해쉬태그할 수 있는 방안을 검토하여 설교 동영상의 방문자 수를 확산시키는 방안을 강구할 수 있다.

주제어 : 토픽모델, 기계학습, 자연어처리, 설교 동영상

I. 서 론

오늘날 소셜미디어 데이터는 숨겨진 비즈니스 기회를 위한 새로운 금광과 가치가 잠재해 있는 원천으로 여겨지고 있다. 기업들은 이러한 방대한 양의 소셜미디어 데이터에 투자하여 어떤 고객 행위와 행동이 더 많은 가치를 창출하고 있는지 규명하기를 시작했다. 즉 많은 기업들이 소셜미디어 데이터로부터 가치를 생성하려고 노력하고 있다 (Kiron, et al., 2014). 그러나 소셜미디어의 가치가 무엇이고 소셜미디어 데이터에서 가치를 창출하는 방법을 찾는다는 것은 매우 힘든 작업이다. 그래서 요즈음 소셜미디어를 인간과 비인간의 상호작용관점에서 가치를 공동 창출하는 매개자

역할에 중점을 둔 소셜미디어 분석 방법들이 소개되고 있다. 사람과 소셜미디어간 상호작용을 통해 생성되는 데이터는 기업과 사용자(사람)에게 동시에 가치를 제공해 준다 (Xie, et al, 2016). 소셜미디어에서 이러한 데이터는 8개 계층에서 나타나고 각 단계에 따라 다른 분석 기법과 도구를 사용한다. 8개 계층은 네트워크, 텍스트, 행위, 하이퍼링크, 위치, 검색엔진, 멀티미디어, 모바일으로 이루어지고, 각 계층에 따른 소셜미디어 분석 유형은 기술, 진단, 예측 그리고 처방적 분석이 있다. 본 연구는 유튜브 설교 동영상 리뷰 분석으로 소셜미디어의 텍스트 계층에서 기술 분석에 속한다고 볼 수 있다.

이와 같은 분석은 대량의 텍스트 데이터에서 패턴을 찾아 마케팅 인사이트 정보을 제공해 준다. 이런 분석 방법은 데이터 마이닝에 기반을 둔 텍스트 분석방법으로 1940 년대 초반부터 기계번역, 정보 검색, 의미 및 감성 분석 및 토픽 모델링의 맥락에서 오랜 기간 연구되었다. 오늘날 이 방법은 다양한 대용량의 데이터를 처리하는 컴퓨터 기술의 발달로 인문사회 분야에서도 쉽게 적용이 가능하다. 본 연구는 토픽 모델링을 적용하여 텍스트를 토픽으로 분해하는 소비자 리뷰같은 구조화되지 않은 텍스트 데이터의 분석이다. 일반적으로 가장 많이 사용하는 토픽 모델링 방법은 Blei et al. (2003)가 개발한 LDA (Latent Dirichlet Allocation)이디.

본 연구는 이 방법을 이용하여 유튜브 설교 동영상에 대한 리뷰(댓글)의 텍스트 데이터를 분석하여 잠재되어 있는 몇 개의 의미 있는 토픽을 찾아내어 리뷰들의 체계적 관리 방안의 제시는 물론 설교 동영상의 확산 방안을 제공하는데 있다. 본 연구의 구성은 첫째, 비지도학습기반의 토픽 모델링을 설교 동영상의 비정형의 리뷰 텍스트 분석에 적용할 수 있는 이론적 배경을 제시한다. 이론적 배경은 소셜미디어 분석에 대한 개요와 기계학습, 토픽모델링 방법을 포함한다. 둘째, 유튜브 설교 동영상 리뷰에서 단어들을 기반으로 토픽을 구성하는 방법을 제시한다. 마지막으로 본 연구의 결론과 한계점 그리고 향후 연구방향을 제시한다.

Ⅱ. 이론적 배경

1 소셜미디어분석

오늘날 이러한 빅데이터 분석의 부상으로 기업 조직에서 실무자와 연구자 모두에게 중요한 연구 분야로 떠오르는 것은 소셜미디어 분석학(Social Media Analytics, SMA)이다. Wedel & Kannan(2016)에 의하면, 소셜미디어 분석학은 사용자들이 생성하여 남긴 데이터(footprint)을 추출하여 미래의 추세나 행동을 예측하는 것을 말한다. 소셜미디어의 8개 계층, 즉 네트워크, 텍스트, 행위, 하이퍼링크, 위치, 검색엔진, 멀티미디어, 모바일의 각 계층에 따른 기술, 진단, 예측 그리고 처방적 분석이 있다. Lee (2018)는 상용 소셜미디어 분석 도구를 사용한 소셜미디어 분석 영역을 감성분석, 소셜네트워크 분석, 통계적 분석, 이미지 및 비디오 분석으로 나누어 소셜미디어 분석 프로세스를 4단계로 나누어 제시하였다. 기업이 소셜미디어 분석 적용 시 소셜미디어 지표 선택의 중요성을 제시하였다.

지금까지 소셜미디어의 댓글에 대한 연구는 UGT관점에서 댓글을 분류하는 유형으로 연구되어 왔다. 페이북, 트위터 등의 댓글에 대한 분류는 사용자의 효용가치에 따라서 정보형, 쾌락형 등의

댓글 혹은 게시글의 수동적인 코딩에 의한 내용분석으로 범주화로 분류하는 연구들이었다 (표 1 참조). Tofesse & Wien (2017)은 소셜미디어 게시물(포스팅)을 내용분석을 통해 감성적 브랜드 게시물, 기능적 브랜드 게시물 등 12개로 분류하였다. Garcia, et al. (2013)은 마케팅에서는 동영상 브랜드와 제품에 대한 광고 반응을 분석하기 위해 댓글의 내용분석을 하였다. 데이터 마이닝 기술의 발달로 최근에서야 대량의 비구조화된 텍스트 데이터에 대해 잠재된 토픽을 찾아내는 연구는 최근에서다. 다음 표는 지금까지 주요 댓글(리뷰)와 게시물(포스팅)를 사용해 수동적 코딩을 통한 내용분석한 연구들을 요약하였다.

[표 1] 소비자 리뷰의 내용분석 연구들의 예

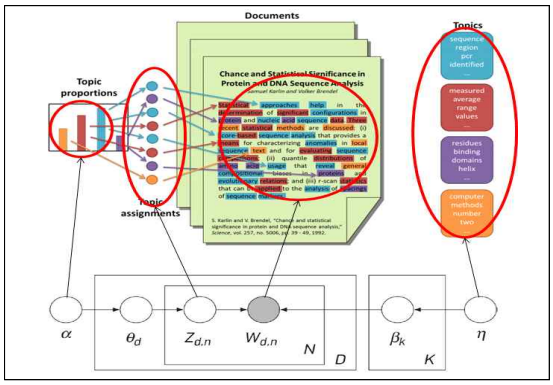
저자(년도)	소셜미디어 (온라인)	대상/방법	분류목적
Berlanga, et all. (2013)	Facebook	posts	설득전략(Ethos, Pathos, Logos)
Duan, et al (2013)	워싱턴 주변 86개 호텔	리뷰	서비스 품질 차원별 감성분석
Luarn et al. (2015)	Facebook Brand page	Posts	posts의 내용을 정보, 오락, 혜택, 사회성으로 분류
Grant et al. (2015)	Youtube 동영상	Youtube Comments	광고와 브랜드에 의한 소비자 태도에 의해 4개로 분류
Gutierrez-Cillan (2017)	Facebook Brand fan page	Brand posts/탐색연구	정보, 이미지, 상호작용 포스트로 분류
Tafsse & Wien (2017)	Facebook Brand fan page	Brand Posts/내용분석	12개로 분류
Jayasingh & Arunkumar (2018)	Facebook Brand Page (Insurance)	Posts/U&G 이론	posts를 정보. 오락, 혜택으로 분류
Rashield & Zeeshan(2018)	Youtube 스마트폰 광고 및 브랜드 동영상	Youtube Comments/내용분석	소비자 반응조사
Lee (2018)	시카고 50고객 130 리뷰 (레스토랑)	리뷰 (Nvivo)	리뷰점수와 각 토픽별 감성점수와의 관계

2 LDA (Latent Dirichlet Allocation)

LDA 모델은 근원 원리를 기반으로 다음 [그림 2]과 같이 다이어그램으로 시각화할 수 있다. 그림은 문서 집합 D와 단어 집합 N 에서 관측 가능한 변수 (모든 단어) Wd, n 로 표현된다. K는 일반적으로 연구자가 설정한 토픽의 수이다. 여기에 많은 잠재 변수를 포함하고 있다. 예를 들어, 특정 단어 문서 d의 n번째 단어 (Wd, n)는 문서 d의 n번째 단어의 토픽(Zd, n)과 토픽별 단어 w의 생성 확률 (βk)의 영향을 받고, 문서 d의 n번째 단어의 토픽(Zd, n)은 문서별 토픽 분포인 Θd에 영향을 받는다 (Blei 2012).

확률적 토픽 모델은 텍스트 코퍼스의 잠재적 토픽 표현을 제공한다. 이 표현은 각 토픽은 단어 등장 분포이며, 각 문서는 토픽의 분포로 모델링한다. 따라서 모형의 적합도 평가와 토픽을 해석

하는 기술이 중요하다. 잠재적 토픽의 의미를 해석하는 방법으로 알려져 있는 것은 없지만, 발견한 토픽이 실제로 의미를 지니고 있는지 그리고 그 의미가 무엇인지 찾는데 주의를 기울여야한다. 토픽 해석과 관련된 중요한 측면은 토픽의 단어 중요도 및 명시적 토픽 이름을 정의하는 것이다. 모델 적합성의 질은 수집된 데이터에 대한 확률적 모델 적합성 계산의 관점에서, 수동 또는 자동 평가를 사용하여 모델 결과의 일관성 및 해석 가능성을 평가한다.



[그림 2] 확률적 토픽 생성 과정

의미 있는 토픽을 만들기 위해 모델을 학습하고 타당성을 검토한 후 토픽 모델의 시각화가 가능하다. 확률적 토픽 모델, 단어와 토픽 행렬, 그리고 문서와 토픽 행력 구조는 토픽 모델을 시각화하기 위한 기반을 형성한다 ([그림 1]참조). 이 시각화는 토픽과 단어 분포, 단어 분포 유사성을 기반으로 한 토픽과 관련된 토픽들, 단일 토픽과 관련된 문서들, 단일 문서와 관련된 토픽들 그리고 단일 단어와 관련된 토픽들을 검색하는 효과적 방법이다. 다른 시각화 아이디어는 토픽의 단어 확률이나 문서의 토픽 확률을 수평 막대그래프나 파이 차트 등으로 표현하는 것이다. 여러 토픽의 상호 작용 대신 단일 토픽의 시각화만 고려할 때 워드클라우드가 기본적인 방법이다. 이 워드클라우드는 텍스트 코퍼스나 토픽 모델을 시각화할 때 자주 사용된다(Gardner et al., 2010; Ramage et al., 2010; Jockers and Mimno, 2013). 이는 하나의 토픽에 다수의 중요한 단어가 선정되어 각 단어의 폰트 크기가 빈도와 같은 단어의 중요도에 비례하도록 그림을 그린다. 단어는 종종 서로 잘 맞도록 배치되어 큰 글씨, 즉 중요도가 있는 단어는 중앙에 배치한다. 이러한 토픽 모델의 시각화는 결과와 토픽의 의미를 해석하는 것뿐만 아니라 두 개의 서로 다른 토픽 모델의 성과를 시각적으로 비교할 수 있는 방법을 제공해 준다.

최근의 소셜미디어의 리뷰(댓글)을 이용하여 문서에서 토픽을 자동적으로 찾아내는 토픽 모델링을 연구한 예를 [표 2]에 정리하였다.

[표 2] 소비자 리뷰의 토픽 모델링 연구들의 예

저자(년도)	소셜미디어 (온라인)	토픽모델/토픽수	목적

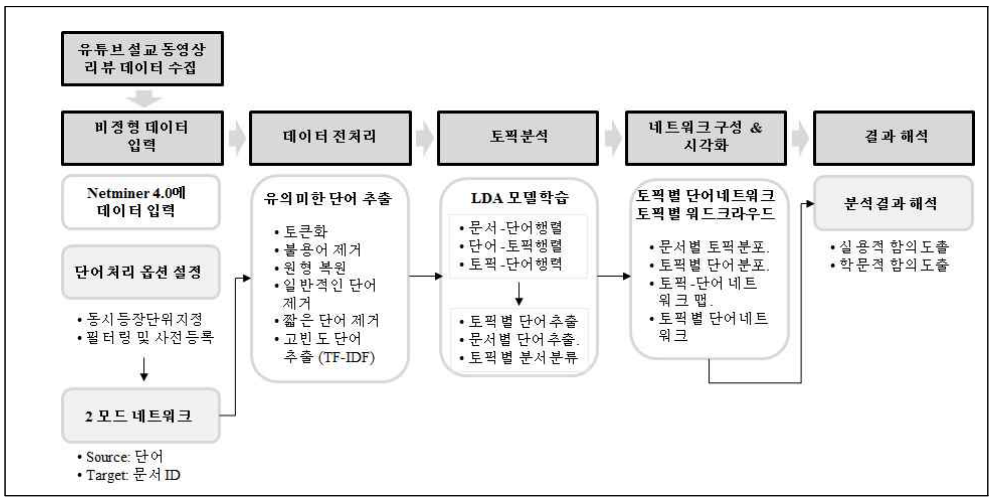
저자(년도)	소셜미디어 (온라인)	토픽모델/토픽수	목적
Hu, et al. (2019)	뉴욕시티의 27,864 호텔 댓글 (TripAdvisor.com)	STM (10)	고객불만 측정
Zhao, et al. (2019)	San Francisco 155개 호텔 127,629 개 리뷰 (TripAdvisor.com)		고객만족
Korfiatis, et al. (2019)	항공사 376,519 승객의 557,208 개 리뷰 (TripAdvisor.com)	STM	서비스 품질
Bastani (2019)	소비자 금융 불만 보호국 (CFPB)	LDA (40)	소비자 불만
Ibrahim, et al. (2019)	386,379개 트윗	LDA (72)/Mallet	온라인 소매 분석
Chhabra & Prajapati (2019)	Amazon Cannon Camera 리뷰	감성분석	의사결정
Sari et al. (2018)	오라인 고객 리뷰 (Tokopedia)	감성분석	e-서비스 품질 특정
김광국 외 (2018)	68,739개 리뷰 모바일 쇼핑 애플리케이션 TOP 20	LDA (30)	모바일 쇼핑 앱 고객만족도
Thelwall (2018)	유튜브 댄스 동영상 댓글	감성분석	성별 감성차이
Ren, et al. (2018)	온라인 리뷰	감성분석	온라인 리뷰와 매출
Boo and Busser (2018)	173개 호텔의 696개 온라인 리뷰	감성분석	호텔평가
Guo et al. (2017)	25,670개의 266,544개 온라인 리뷰	LDA (30)	고객만족 측정
Chang, et al. (2017)	Hillton Hotel (TripAdvisor)	SVM 감성분석	
Moen, et al. 2017	30개 제품 카테고리 1,147,488 개 리뷰	자연어 처리	리뷰와 매출간 관계
Hakh, et al (2017)	미국 6개 항공사의 14,640 트윗	감성분석	감성분석 기법 비교
Ernst et al (2017)	연설 동영상에 대한 유튜브 커멘트	토픽 및 감성분석	

Ⅲ. 실증 분석

본 연구절차는 [그림 3]과 같이 6단계로 구성된다. 우선 유튜브 설교 동영상 리뷰 데이터의 추출이다. 본 연구는 Neminer 4.0의 Youtube Collector를 사용하여 추출하였다. B교회 목사님의 유튜브 설교 동영상에 대한 댓글을 2019년 1월1일부터 7월30일까지 사용자 리뷰와 답글을 포함하여 361개를 추출하였다.

둘째는 수집된 유튜브 리뷰 데이터인 텍스트 데이터를 Netminer 4.0에 불러들이고 단어가 동시등장 (Co-occurrence)한 명사에 의해 분석하였다. 이와 동시에 스팸문자, 특수문자, URL 등을 제거하는 필터링과 설교 유튜브 동영상 리뷰 내에서 유의어 사전, 동의어 사전, 확정어 사전, 스팸사전을 구축하여 등록하였다. 처음 361개 문서로 이 과정을 거친 후 342개 문서와 3,060단어를 추출하였다. 세 번째 단계로 유의미한 단어를 추출하는 데이터 전처리 작업이다. 전처리를 위해 토큰화, 불용어 제거, 원형 추출을 Netminer 4.0의 텍스트 마이닝에 의해 수행하였다. 그리고 한 글

짜로 된 단어와 일상적으로 사용하고 있는 단어를 제거한 후 TF-IDF를 계산하여 0.5이상의 고빈도의 2,393개 단어가 추출되었다.



[그림 3] 토픽 모델링 프로세스

이제야 비로소 토픽 모델링에 해당하는 네 번째 단계인 LDA에 의한 비지도 학습(Unsupervised Learning)이다. 텍스트 데이터에 LDA를 적용할 때, 연구자는 생성될 토픽의 수를 선택해야 한다(Maier et al., 2018). Westerlund et al. (2018)의 개념에 따라, 본 연구는 중첩을 피하고 토픽의 해석 가능성을 보장하는 시행착오를 토대로 토픽의 수를 선택했다. Calheiros et al. (2017)에 따르면 소수의 토픽은 토픽간 중복을 피하는 데 도움이 된다. 이것은 차례로 구조 타당성을 확보하는 데 도움이 된다 (Schmiedel et al., 2018). 또한, Lei and Law (2015)는 소수의 토픽은 일반적으로 소비자 리뷰에서 핵심 속성을 포함하고 있음을 제시하고 있다. 이 개념은 Maier et al. (2018)의 주장과 잘 부합되며, 가능한 토픽의 수는 이론적인 개념, 연구 맥락 및 연구의 목적과 일치하도록 세분화하는 것이 실용적이다. 본 연구는 최적 토픽 수를 시행착오를 통해 토픽이 서로 배타적으로 나누어지고, 상위 10까지 등장하는 단어들이 서로 겹쳐지지 않을 때를 선택했다. 또한 토픽-단어의 2 모드 네트워크로 표현했을 때 스타형일 때 가장 설명력이 높은 것으로 판단하여, 본 연구는 토픽 수를 5개로 선택하였다.

토픽 수 K와 더불어 문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터 α와 토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터 β도 결정되어야 잠재적 토픽이 문서들에서 결정된다. 알파 값을 높게 잡으면 문서의 토픽 연관확률이 동일하게 나타나며, 이는 문서에 모든 토픽이 동일한 확률로 분포함을 의미한다. 베타값을 높게 잡을 시는 토픽의 단어 연관확률이 동일하며, 토픽에 대한 모든 단어의 연관확률이 동일함을 의미한다. 따라서 이들 값의 최적 값은 없으나 본 연구는 기존 연구자의 결과를 기반으로 설정하였다. Griffiths

and Steyvers (2004)에 따르면 α와 β의 선택은 LDA 결과에 중요한 영향을 줄 수 있다. 그들은 α=50/k (k: 토픽의 수)와 β=0.1을 사용했다. β에 작은 값을 사용한 이유는 연구의 특정 영역을 나타내는데 더 많은 토픽을 만들기 위해서였다. Steyvers와 Griffiths (2007)는 파라미터 α와 β가 토픽의 수와 어휘 크기에 따라 다르다는 것을 입증했다. 또 그들은 β=0.01 이외의 값을 그리고 α=50/k 값을 사용할 것을 권고하고 있다. 최근에 Zhao, et al. (2015)은 LDA의 파라미터 α 값은 0.01과 0.1 사이, β 값은 0.01일 때 가장 설명력이 높은 토픽 모델을 만들 수 있음을 제안했다.

[표 3] LDA의 파라메터 관련 연구 (Marwa Naili, el al., 2017)

저자	응용분야	테스트 코퍼스	α	β
Steyvers & Griffiths (2007)	토픽 분류	TASA	50/K	0.01
Lu et. al., (2011)	토픽 분류 및 정보검색	TDT2	[0.1, 0.5] or [0.5, 2]	0.01
Zhau et al.(2015)	토픽 식별	Reuters-21578	[0.01, 0.1]	0.01
Naili, el al. (2017)	토픽 식별	AI-Watan	0.5	0.01

본 연구에서 α=50/k (k: 토픽의 수)에 의해 10으로 하였으며, β=0.01을 사용했다(Griffiths & 2004, 2007). 관찰된 문서를 통해 잠재된 토픽을 추출하는데 기반이 되는 문서-단어 행렬의 2 모드 네트워크 (source: 단어, target: 문서 ID)가 추출되었다. 이 행렬은 문서 (열) 내의 단어 (행)의 분포 (빈도)를 나타낸다. 실제로 텍스트 마이닝에서 단어-문서 행렬은 토픽 모델링의 LDA 알고리즘에 대한 입력으로 사용되는 벡터-공간 표현의 2차원 매트릭스 구조를 생성한다. 행렬의 각 셀은 문서에 단어가 등장하는 빈도를 표현한다. 단어-문서 행렬은 빈도를 통해 문서와 문서 간의 관계를 특징짓는다(Delen & Crossland, 2008).

토픽수 5, α는 10, β는 0.01에 의해 다음과 같은 토픽-단어 행렬, 문서-토픽 행렬을 얻었다. 다음 [표 4]와 [표 5]는 토픽별 단어의 연관확률과 토픽별 문서의 연관확률을 각각 보여 주고 있다. 이는 최초 LDA 분석 결과 토픽별 단어의 연관확률(영향력)을 표시한다. 각 토픽별로 연관확률이 가장 높은 단어를 1st Keyword, 두 번째 단어를 2nd Keyword 순으로 나타냈다. 또 최초 LDA 분석 결과 문서별 토픽의 연관확률(영향력)을 표시한다. 각 문서별로 연관확률이 가장 높은 토픽을 1st topic, 두 번째 단어를 2nd topic 순으로 하였다.

LDA를 이용한 토픽 모델링에서 비지도학습을 위해 가장 널리 사용하는 샘플링 방법이 깁스 샘플링(**Gibbs sampling**)이다. Gibbs 샘플링은 Monte Carlo Markov-chain (MCMC) 알고리즘으로, 각 변수의 조건부 분포를 효율적으로 계산할 수 있는 경우 결합 분포에서 샘플을 생성하는 통계적 추론의 강력한 기법이다. 본 연구는 이 깁스 샘플링 방법을 사용하여 1000회 반복을 통해 학습하였다.

[표 4] 토픽별 단어의 연관확률의 예

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
생각	0.000158	0.000419	0.037395	0.000209	0.025286
성도	0.003374	0.004165	0.024841	0.000151	0.000180
신영복	0.005432	0.000492	0.020009	0.002862	0.013545
동성애	0.016644	0.000049	0.019700	0.000130	0.000180
공산주의	0.000716	0.001206	0.018544	0.000050	0.006969
예수	0.002071	0.000373	0.015509	0.047274	0.000340
종교	0.001522	0.000033	0.013297	0.000686	0.000043
아멘	0.005586	0.001669	0.013080	0.000772	0.002546
인간	0.001307	0.000186	0.013002	0.000038	0.006142
좌파	0.000204	0.000175	0.012696	0.000166	0.026503

[표 5] 문서-토픽 연관확률의 예

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Doc. 008	0.083179	0.095574	0.102305	0.649413	0.069529
Doc. 007	0.083247	0.095687	0.102430	0.648987	0.069648
Doc. 040	0.126759	0.454237	0.128703	0.159898	0.130403
Doc. 051	0.130677	0.148241	0.177765	0.382275	0.161041
Doc. 100	0.139242	0.330980	0.154150	0.140529	0.235098
Doc. 259	0.168331	0.183332	0.199150	0.191199	0.257989
Doc. 032	0.170675	0.222639	0.227349	0.159066	0.220271
Doc. 028	0.171497	0.182413	0.207194	0.199534	0.239362
Doc. 013	0.175105	0.236135	0.204564	0.190031	0.194166
Doc. 097	0.175870	0.175213	0.278873	0.177656	0.192389

[표 6]와 같이 정리하면 첫 번째 토픽은 감사, 은혜, 성령, 그리스도, 아버지 등으로, 두 번째 토픽은 하나님, 사랑, 기독교, 마음, 사랑으로, 세 번째 토픽은 생각, 성도, 신영복, 동성애, 공산주의로, 네 번째 토픽은 감사, 예수, 설교, 교인, 이름 등으로, 마지막 5번째 토픽은 기도, 존경, 설교, 좌파, 나라 등으로 구성되어 있음을 알 수 있다. 토픽별 문서의 수는 첫 번째 토픽이 62개 문서, 두 번째 토픽은 71개 문서, 토픽 3은 78개 문서, 토픽 4는 36개 문서, 토픽 5는 95개 문서에 나타나고 있음을 보여 주었다. 동성애에 대한 의견 표현으로 설교 동영상 댓글에서 주제 5와 3에 부정적으로 비취는 댓글들이 나타났음을 볼 수 있다.

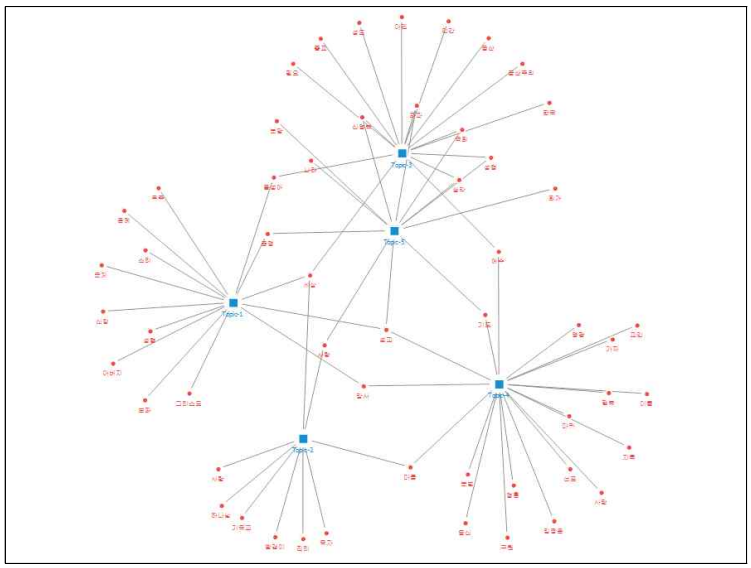
[표 6] 토픽 분석 요약

	문서의 수	1위 단어	2위 단어	3위 단어	4위 단어	5위 단어
Topic 1	62	감사	은혜	성령	그리스도	아버지
Topic 2	71	하나님	사람	기독교	마음	사랑
Topic 3	78	생각	성도	신영복	동성애	공산주의
Topic 4	36	감사	예수	설교	교인	이름
Topic 5	95	기도	존경	설교	좌파	나라

마지막으로 시각화 작업이다. 다음 그림은 전 처리작업 후 워드클라우드를 보여 주고 있다. 이 그림에 의하면, 존경, 기도, 말, 생각, 예수 등이 가장 높은 빈도를 보여 주고 있음을 알 수 있다.

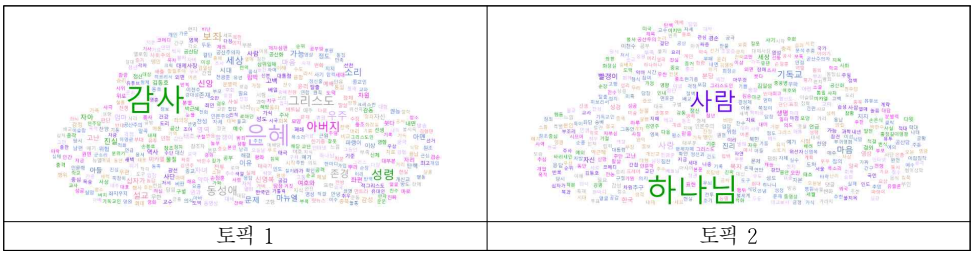
[그림 4] 전처리 작업 후의 워드클라우드

다음의 시각화는 [그림 5]와 같이 토픽-단어 네트워크이다. 토픽-단어 네트워크는 모든 토픽과 모든 단어간의 2-모드 네트워크로 모든 단어의 토픽 연관 확률이 0이상이므로 (디클레이분포) 즉 단어-토픽간의 링크의 가중치는 단어-토픽의 연관확률이고 이는 모두 0이상으로 링크가 존재한다. 따라서 5개 토픽과 모든 단어간 네트워크의 표현은 모든 단어와 토픽에 연결이 표현되므로 식별이 어렵다. 따라서 본 연구는 토픽별 연관확률이 낮은 경우는 제외하고 토픽별 영향력이 높은 주요단어만 추출하여 네트워크 분석 토픽별 상위 70위에 있는 단어들로 구성하여 단어-토픽 네트워크는 다음 [그림 5]와 같이 표현하였다. 토픽간 중첩된 단어에 의해 토픽간 연관관계를 보여주고 있다.



[그림 5] 토픽-단어의 2모드 네트워크

이 네트워크로는 토픽별 단어의 등장 빈도를 알 수가 없어 다음 [그림 6]은 토픽 내 등장하는 단어의 빈도에 의한 워드클라우드를 보여 준다. 토픽 1과 토픽 2에 대해서만 에로 제시하였다. 토픽 1은 감사와 은혜가 높은 등장 빈도를, 토픽 2는 하나님과 사람이, 토픽 3은 생각, 성도가, 토픽 4는 예수, 감사가, 마지막 토픽 5는 존경과 기도가 많이 등장하고 있음을 보여 준다.



토픽 1	토픽 2

[그림 6] 토픽 1과 토픽 2 워드클라우드

이 토픽 5개 중 3개 토픽 1, 2, 4는 온라인 고객 리뷰들처럼 본 설교 동영상의 핵심 속성으로 볼 수 있지만, 다른 두 토픽 3과 5는 설교 동영상과는 거리가 있는 리뷰가 포함된 것을 볼 수 있다. 이는 설교 동영상보다는 설교자의 이전 말이나 행위들에 대해 이야기한 리뷰들이라고 볼 수 있다. 따라서 교회 입장에 볼 때 리뷰에 대한 체계적이고 객관적인 평가와 관리가 요구된다. 사용자들이 기록한 리뷰들을 토픽별로 분류하여 긍정적인 리뷰들과 가장 관련성이 있는 리뷰들을 부각시키는 리뷰의 재배열 방식이 필요하다. 또한 긍적적인 리뷰들을 포함한 토픽을 사용하여 해쉬태그할 수 있는 방안을 검토하여 설교 동영상의 방문자 수를 확산시키는 방안을 강구할 수 있다.

Ⅳ. 결론 및 향후 연구방향

소셜미디어 확산과 함께 오늘날 교회들도 페이스북, 인스타그램, 유튜브 그리고 모바일 앱등의 소셜미디어를 활용하여 선교활동을 넓히고 있다. 이러한 미디어의 사용은 시간과 공간을 초월하여 선교를 활 수 있는 많은 장점을 제공해 줌과 동시에 불특정 다수의 사람들에 의한 부정적인 공격을 받을 수 있는 기회가 열려있다. 바로 이러한 활동은 소셜미디어 상에 사람들이 남기는 흔적들이다. 이러한 흔적은 동영상, 사진, 텍스트, 이모티 등으로 표현된다. 이러한 데이터들은 정형화보다는 비정형화된 데이터로 컴퓨터의 대용량화 및 처리 속도의 발달로 이러한 데이터를 처리할 수 있게 되었다. 따라서 교회들이 활용하고 있는 소셜미디어의 평가는 물론 활용방안을 찾는 것이 필요하다.

본 연구는 유튜브 설교 동영상의 리뷰를 토픽 모델링으로 설교에 대한 반응을 몇 개의 토픽별로 나누어 분석하였다. 분석결과 5개의 토픽으로 분류하였다. 토픽 5개 중 3개 토픽 1, 2, 4는 온라인 고객 리뷰들처럼 본 설교 동영상의 핵심 속성으로 볼 수 있지만, 다른 두 토픽 3과 5는 설교 동영상과는 거리가 있는 리뷰가 포함된 것을 볼 수 있다. 이는 설교 동영상보다는 설교자의 이전 말이나 행위들에 대해 이야기한 리뷰들이라고 볼 수 있다. 따라서 교회 입장에 볼 때 리뷰에 대한 체계적이고 객관적인 평가와 관리가 요구된다. 사용자들이 기록한 리뷰들을 토필별로 분류하여 긍정적인 리뷰들을 부각시키는 리뷰의 재배열 방식이 필요하다. 또한 긍적적인 리뷰들을 포함한 토픽을 사용하여 해쉬태그할 수 있는 방안을 검토하여 설교 동영상의 방문자 수를 확산시키는 방안을 강구할 수 있다.

본 연구는 이와 같이 실용적 학문적 시사점을 제시해 줄 수 있지만 토픽모델링은 사용자들이 표현한 있는 그대로의 문장과 단어를 기반으로 했기에 숨어있는 사용자들의 의도, 생각, 깊이 있는 의미들을 찾아내는 데는 한계가 있다. 오늘날 이를 해소하는 방안으로 감성분석과 결합한 연구가 요구된다.

참고문헌

김광국 외 (2018). 사용자 리뷰 토픽분석을 활용한 모바일 쇼핑 앱 고객만족도에 관한 연구, The Journal of Society for e-Business Studies, Vol.23, No.4, pp.41-62.

문동지, 연다인, 김희웅 (2018). 토픽 모델링 기반 한국 노인의 행복과 불행 이슈 분석. Information systems review, 20(2), 139-161.

박은준, 김영지, 박찬숙 (2017). 텍스트네트워크분석을 활용한 국내 · 외 호스피스 간호 연구 주제의 비교 분석. Journal of Korean Academy of Nursing, 47(5), 600-612.

윤지은, 서창진 (2018). 토픽모델링과 에고 네트워크 분석을 활용한 스마트 헬스케어 연구동향 분석. 한국디지털콘텐츠학회논문지, 19(5), 981-993.

임병학 (2019). 사회 연결망 분석을 이용한 성경 네트워크 분석에 관한 연구. 로고스경영연구, 13(4), 109-124.

조성배, 신신애, 강동석 (2018). 토픽 모델링을 이용한 개방형 혁신 연구동향 분석 및 정책 방향 모색. 정보화 정책, 25(3), 52-74.

최종산 (2017). 소셜 빅데이터를 이용한 외국인의 한식에 대한 인식 분석. 예술인문사회융합멀티미디어논문지, 7(8), 427-437.

최영출 (2012). 사회적기업의 정책요소분석을 통한 적정모형 탐색. 한국비교정부학보, 16(1), 149-166.

Blei, D. M. (2012). “Probabilistic topic models,” Communications of the ACM, Vol 55, No. 4, pp. 77-84.

Blei, D., Ng, A., and Jordan, M.(2003). “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol. 3, pp. 993-1022.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan). pp. 993 - 1022.

Boo and Busser (2018).

Calheiros AC, Moro S, Rita P (2017). Sentiment classification of consumer-generated online reviews using topic modeling. Journal of Hospital Market Management, 26(7), pp. 675 - 693.

Cao D, Ji R, Lin D, Li S (2014). Visual sentiment topic model based microblog image sentiment analysis. Multimed Tools. Applicsation, 2016(75), pp.8955 - 8968.

Chang, W., Wang, J. (2018). Mine is yours? Using sentiment analysis to explore the degree of risk in the sharing economy. Electronic commerce research and application, 28. pp.141-158.

Chhabra & Prajapati (2018). Sentiment Analysis of Amazon Canon Camera Review using Hybrid Method. International Journal of Computer Applications, 82(5),

Do T-M-T, Gatica-Perez D (2010). By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: Proceedings of the 9th international conference on mobile and ubiquitous multimedia (MUM'10), 1 - 3

Duan W, Cao Q, Yu Y, Levy S (2013) Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In: Proceedings of the 46th hawaii international conference on system sciences, pp 3119 - 3128.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59,

467 - 483.
Hu, Nan, Ting Zhang, Baojun Gao, Indranil Bose (2019). What do hotel customers complain about? Text analysis using structural topic model. Tourism Management, 72. pp. 417-426.
Ibrahim, N.F., Wang, X.(2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. Decision Support Systems, 121, pp. 37-50.
Khan (2018). Creating Value with Social Media Analytics: managing, Aligning, and Mining Sopcial Media Text, Networks, Action, Locations. Apps, Hyperlinks. Multimedia and Search Engines Data, 2018.
Korfiatis, N., Stamolampros, P., et. al., (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Expert Systems with Applications. 116, pp.472-486.
Lai & To (2015). Social Media Content Analysis: A Grounded Approach, Journal Of Electronic Commerce Research, VOL 16, NO 2.
Lai, L.S.L. And E. Turban (2008). "Groups Formation And Operations In The Web 2.0 Environment And Social Networks," Group Decision And Negotiation, Vol. 17, No. 5:387-402.
Lee, I., (2018). Social media analytics for enterprises: Typology, methods, and processes.
Lu B, Ott M, Cardi C, Tsou BK (2011). Multi-aspect sentiment analysis with topic models. 11th IEEE International Conference on Data Mining Workshops, pp 1 - 8
Marwa Naili, Anja Chaibi, Henda Ghézala (2017), Arabic topic identification based on empirical studies of topic models. INRIA, Volume 27.
Moen, Havro, L., Bjering, E. (2017). Online consumers reviews: Examining the moderating effects of product type and product popularity on the review impact on sales. Cogent Business & Management, 4. 1-20.
Reisenbichler, M. & Reutterer, T.(2019). Topic modeling in marketing: recent advances and research opportunities. Journal of Business Economics, 89, pp.327 - 356.
Ren, J., Yeoh, W., Ee, M., Popovic, A. (2018). Online Consumer Reviews and Sales: Examining the Chicken-Egg Relationships. JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY, 69(3):449 - 460
Schmiedel, T., Oliver Muller, and Jan vom Brocke (2018). Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture. Organizational Research Methods, pp.1-28.
Steyvers, M., Griffiths, T. 920070. Probabilistic topic models. In Handbook of Latent Semantic Analysis; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, USA, pp. 424 - 440.
Thelwall, M. (2018). Social media analytics for YouTube comments: potential and limitations. InternatIonal Journal of Social Research Methodology, 21(3), pp.303 - 316
Turban, E., J. Strauss, And L.S.L. Lai (2015). Social Commerce - An IS And Marketing Perspective, Springer.
Wedel & Kannan (2016), Marketing Analytics For Data-Rich Environments, Journal Of Marketing: AMA/MSI Special Issue, Vol. 80 (November 2016), 97-121.
Westerlund, M., Leminen, S., & Rajahonka, M. (2018). A Topic Modelling Analysis of Living Labs Research. Technology Innovation Management Review, 8(7): 40-51.