[[1]](#footnote-1)

주제 모델을 이용한 구조화 및 비구조화 문서에서 관찰되지 않은 변수의 추정 평가

콜린N.우다노르,오니니에E.뉴케,IEEE회원대학원생,*그리고조지E.*오케레케

***초록—효과적인 데이터 수집을 위해 연구자들은 종종 어디, 무엇, 어떻게라는 세 가지*도전에 직면합니다. 연구 가능한 데이터를 어디에서 찾을 수 있는지, 어떤 도구와방법론을 사용하여 웹사이트에서 데이터를 수집할 것인지, 그리고 필요한 분석을 수행하고 유의미한지식을 추출하는 방법은 무엇인지에 대한 질문입니다. 본 연구는 특히 머신러닝 및인공지능 분야에서 사용자 트윗이 연구 방향에 미칠 수 있는 영향의 가능성과범위를 조사합니다. 본 논문에서는 20개의 인공지능 및 머신러닝 관련 계정에서 수집한35,860개의 비정형 데이터(트윗)에서 LDA(Latent Dirichlet Allocation) 주제 모델링 기법을 사용하여 머신러닝연구 주제의 인기를 발견하고, 42년간의 Neural Information Processing Systems(NIPS) 회의 논문데이터셋인 7,241개의 정형 문서를 대조군으로 사용합니다. LSI(Latent Semantic Index)와 HDP(Hierarchical DirichletProcess)를 사용하여 LDA의 성능을 비교합니다. Bag of Words 및 TF-IDF(term frequency-inversedocument frequency)와 같은 임베딩 방법을 사용하여 코퍼스를 인코딩하고 비교합니다. 결과는 정형데이터셋을 사용하면 더 나은 분류가 보장된다는 것을 시사하지만, 비정형 데이터셋도 quiteinformative합니다. 그러나 t-test 결과 두 데이터셋의 결과 차이는 유의하지 않았습니다. LDA모델은 주제별로 LSI 및 HDP를 일관되게 능가했습니다. Gensim 및 Mallet Python프레임워크 비교 결과, Mallet이 Gensim보다 더 나은 주제 모델링 결과를 약속했습니다.**

***주요*용어—Bagofwords,Gensim,LatentDirichletAllocation,Machinelearning,Mallet,Topic models, Twitter.**

# I.서론

사회적 미디어가 우리의 일상, 여가, 경제, 정치, 연구 생활의 모든 분야에 영향을 미쳤음은의심할 여지가 없습니다. 또한 사회적 미디어는 연구 데이터의 생성과 관리 수단으로도 자리잡았습니다. 사이버공간에서 사회적 미디어를 통해 사람들과의 상호작용이나 전자상거래 사이트, 온라인 게임, 스포츠 등에서 기계와의상호작용으로 인해 enormous한 양의 데이터가 폭발적으로 증가하고 있습니다. 연구자들은 이러한 새로운 데이터 고속도로를활용하여 인간 행동, 구매 패턴, 연구 동향, 의견 조사, 기타 결과 등을 연구하고,개인의 건강 상태, 지진 및 기타 자연 재해의 조기 경고 신호 등을 모니터링하고 있습니다.

트위터의 빠른 콘텐츠 전달, 확산, 그리고 검색 및 스트리밍 애플리케이션 프로그래머 인터페이스(Search and Stream APIs)의 제공은 트위터를 연구 데이터 수집의1위 위치로 올려놓았습니다. 이는 페이스북과 같은 다른 소셜 미디어 사이트보다 더 높은 위치입니다. 페이스북과 달리 트위터의 계정 중 10% 미만이비공개 계정으로, 연구자들이 90% 이상의 공개 트위터 데이터를 수집하는 데 도움이 됩니다[1]. 트위터는 문자 수 제한과 빠른 콘텐츠 전달로 인해소셜 미디어의 "SMS"로 여겨집니다. 트위터의 학계에서의 인기는 리트윗[2-4]을 기반으로 연구자의 영향을 측정하기 위한 대체 지표(altmetric)인 t-인자 개발에 대한 제안으로 이어졌습니다.

트위터가 학술지 및 연구 논문에 미치는 영향에 대해 본만과 하운슈일드[2]가 분석한 바, 트윗을 통한 논문 확산과인용 영향력을 고려했습니다. 연구 결과, 트위터 계정이 있는 학술지는 소셜 미디어(트위터)의 존재가 없는 학술지에 비해 인용횟수가 더 많았습니다. 연구자들은 학술지가 트위터에 참여하는 최적의 방법을 찾기 위해 노력했습니다. 트위터 계정 지표인 총트윗 수, 팔로워 수, 팔로잉 수 등을 고려하여 인용에 미치는 영향력을 분석했습니다. 또한, 대체 지표 제공업체인플럼 애널리틱스도 사용되었습니다. 350개 학술지에 게재된 4,176편의 연구 논문 데이터셋에 대해 학생의 T-검정과 회귀 분석이라는 두가지 통계 분석이 적용되었습니다. 학술지는 네 가지 분류로 나뉘었습니다(학술지 계정, 소유주 계정, 출판사 계정, 트위터 계정없음). 이 연구는 트위터 프로필의 팔로워 수가 특정 논문이 받는 트윗 수와 인용 횟수에 크게 영향을 미친다는 결론을 내렸습니다.

트위터에서 출판물의 영향을 측정하기 위한 t-지수를 제안한 연구에서, 인용지수(h-index)로 유명한 연구 성과 측정 지수를 기반으로 한 t-지수를 제안했습니다. 이 연구에서는 Scopus에서 단일 논문을 사용했습니다. 연구자들은 특정 연구자 또는 논문 출판물의 모든 트윗과 리트윗을 분석을 위해 단일 트윗으로 결합하기 위해 h-지수의 공식을 기술로 사용했습니다.

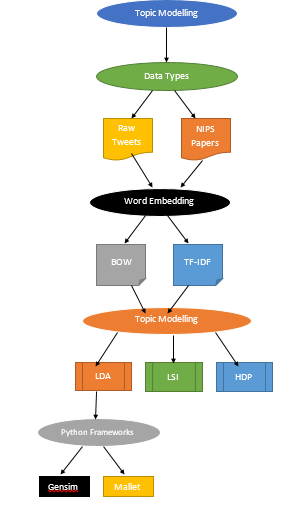
An overview of Twitter activity as the basis of scholarly twitter metricswas given by Haustein [4]. The author analysed 24 million tweets withthe intention of analysing tweets measurable with respect to research evaluation. Theirstudy finds that majority of tweets related to research output are mostlycreated by the academia, hence the impact is more on academic communicationrather than societal. They also observed that the peak of transmission ofthese academic tweets happen shortly after publication. There is however a lowdiscussion of academic tweets generally among twitter users. Tweets, if properly analysedcan reveal trending topics of discussions among Twitter users, and provide insightsto the general perception of the groups involved in the tweet, retweet, and mentions.

주제는 트윗, 문장, 단락, 뉴스 기사 또는 책장에서 단어의 공동 발생 패턴으로 의미론적으로 정의됩니다. 문서는주제와 단어로 분해되고 모델링됩니다. 주제 모델링에는 Latent DirichletAllocation (LDA), Hierarchical Dirichlet Process (HDP), Spherical HDP,Latent Semantic Analysis/Latent Semantic Indexing (LSI), pLSA 등이포함됩니다. Latent Dirichlet Allocation (LDA)은 문서를 특정 비율의주제 모음으로 고려합니다. 이는 2003년에 개발된 단어 모음모델로, 문서 분류, 감정 분석, 바이오인포매틱스 등 여러분야에 적용됩니다. 모델이 관찰하는 특징은 문서에 등장하는 단어입니다.

이 작업은 비정형 및 정형 데이터셋과 주제 모델링 기법을 사용하여 트렌드 있는 머신러닝 및 AI 주제를 발견하고, 이러한 주제가 머신러닝 연구에 미치는 영향을 분석합니다.

이 연구에서는 트위터 마이크로블로그와 NIPS 컨퍼런스(1975-2017년)의 머신러닝 논문에서 수집한 원시 트윗 데이터를 대상으로 문서 모델링을 수행하여 숨겨진 주제를 발견하는 방법을 제시합니다. 이 연구에서는 LDA,LSI, HDP 세 가지 주제 모델링 기법을 사용하여 데이터셋을 모델링하고, 각 기법의 성능을 비교하여 최적의 알고리즘을 결정합니다. LDA 모델은 주제 모델링의 벤치마크로 사용되며, 일반적으로Bag of Words(BOW) 임베딩을 사용합니다. 그러나 본 연구에서는 Bag of Words(BOW) 대신 Term Frequency-Inverse Document Frequency(TF-IDF) 임베딩을 사용할 경우 결과에 유의미한 차이가 있는지 탐구합니다.

Fig. 1 gives a graphical overview of the research processes in this work. Section 2 gives a brief background into topic modelling techniques and how they work. The section also reviews related works done with Twitter data, with multiple topic modelling techniques and concludes with a review of topic modelling techniques. Section 3 lays out the experimental works, showing the two major types of data, tweets and NIPS papers used for the experiment. The data pre-processing and transformation methods used are also described. The experiment with LDA, LSI, and HDP topic modelling are also described, as well as the data embedding techniques, and the Python frameworks used. Section 4 presents four different results from each of the experimental stages, while Section 5 discusses the findings and compares them with the ground truths.



**Fig.**1.이연구의 워크플로우

# II.배경

텍스트 마이닝은 텍스트 코퍼스에서 다양한 주제를 포함할 수 있는 숨겨진 지식을 발견하는과정으로, 하나 이상의 텍스트 마이닝 또는 머신러닝 알고리즘을 사용합니다. 텍스트 마이닝 작업의출력 결과는 용어-문서 행렬(TDM)입니다. TDM은 주제 모델에 대한 입력으로 사용되며, 이는 문서에대한 레이블이 없는 문서에서 자동으로 학습하는 생성 과정입니다. [5]의 저자는 확률적 잠재의미 분석(pLSA)을 사용하여 주제 모델을 제안한 최초의 인물 중 한 명입니다. 가장인기 있는 문서 모델 중 하나는 Latent Dirichlet Allocation(LDA)[6]로, 이는 서로 상관관계가없는 것으로 생각되는 주제를 분석하기 위한 베이지안 혼합 방법입니다. 상관관계가 있는 주제모델(CTM)은 LDA의 확장판으로, 주제 간 상관관계를 허용합니다. 머신러닝과 자연어 처리에서 주제 모델은코퍼스 내 문서의 용어 빈도 발생에 대한 프레임워크를 제공합니다[7]. 주제 모델은 혼합 멤버십 모델이므로, 유니그램 또는 유니그램 모델의 혼합과 다릅니다.

1. *LDA의 작동 원리*

LDA는 문서를 특정 비율의 주제들의 모음으로 간주합니다. 2003년에 개발된 이 모델은 문서 분류, 감정 분석, 생물정보학 등 다양한θ분야에 적용됩니다. LDA가 관찰하는 특징은 문서에 등장하는 단어들입니다. 다른 매개변수들은 잠재적(은닉적)입니다. 예를 들어, 단어에 할당되는 주제가 있습니다. 따라서모든 문서는 이러한 주제의 혼합물입니다. Fig. 2는 모델이 문서, 주제, 단어라는 세 가지 엔티티를 어떻게 결합하는지 보여줍니다. M은문서의 수, N은 주제의 수, 그리고 는 각 주제의 분포입니다. 단일 문서에는 N개의 단어가 있으며, 각 단어에는 N개의 주제가 있습니다.

**LDA모델의플레이트**표기법 표현, Fig. 2

각 문서는 포함된 주제와 해당 주제가 차지하는 비율로 설명할 수 있습니다. 예를 들어, 문서에는 AI가 50%, OS가 30%, 게임이 10% 등 포함될 수 있습니다. 모델은 이러한 문서를 받아 특정 주제에서 적절한 수의 단어를 추출하여 새로운 문서를 생성합니다.

LDA는 문서가 다양한 버킷 또는 주제에서 가져온 단어의 모음인 백 오브 워즈모델입니다. 문서에서 생성할 주제의 수를 지정하고, 주제가 어떻게 구성되어야 하는지 규칙을 정해야합니다. 이러한 규칙은 하이퍼파라미터이며, 알파와 베타로 표현됩니다. 이들은 모두 디리클레 분포의 매개변수입니다.

알파는 문서별 주제 분포를 제어하며, 베타는 주제별 단어 분포를 담당합니다. 알파 값이 높을 경우 문서가 대부분의 주제를 포함할 가능성이 높음을 의미하며, 단일 주제만 포함하지는 않음을 의미합니다. 반면, 알파 값이 낮을 경우 문서가 몇 가지 주제만 포함할 가능성이 높음을 의미합니다. 반면, 베타 값이 높을 경우 각 주제가 대부분의 단어를 포함할 가능성이 높음을 의미하며, 베타 값이 낮을 경우 주제가 몇 가지 단어만 포함할 가능성이 높음을 의미합니다. 요약하면, 알파 값이 높을 경우 문서들이 서로 더 유사하게 보이고, 베타 값이 높을 경우 주제들이 서로 유사하게 보입니다. 다음 공식은 지금까지 말한 모든 내용을 포착합니다. 이는 모델의 총 확률로, 방정식 (1)에 의해 주어집니다.

P(W, Z, q, j, a, b) = ) Jj;a)Zj,t;J)P(Wj,t|j) (1)

특정 주제의 수에는 제한이 없습니다. 이는 애플리케이션 및 사용 사례에 따라 달라집니다.

1. **LDA 모델 사양**

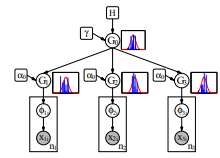
LDA는 앞서 언급한 바와 같이 주제 모델링을 위해 세 단계의 생성 과정을 사용합니다.

1. Term document distribution determined by: ~ Dirichlet(
2. 문서W의주제분포에서주제q의비율은다음과같이 결정됩니다: q ~ Dirichlet (a).
3. 각 N개의 단어Wi에 대해, 다음을수행할 수 있습니다;
4. 주제Zi~다항분포(q)선택
5. 주제에 따라 조건화된다항 확률 분포에서단어를 Wi로 선택합니다.

Zi:P(Wi|Zi,b),여기서b는주제의단어분포를나타내며,또한특정주제에서단어가발생할 확률을 포함합니다.

1. **계층적디리클레과정**(HDP)

HDP는 디리클레 프로세스 혼합 모델을 기반으로**하며, 각 그룹의 디리클레 프로세스는 기본**분포를 공유합니다. HDP는 일반적으로 비모수적인 그룹화된데이터를 클러스터링하는 데 사용됩니다. HDP는 주제모델로 적용될 수 있으며, 단어는 문서로구성되고 각 문서는 단어의 집합으로 이루어집니다.HDP의 상태 수는 제한되지 않으며 데이터에서학습될 수 있습니다. 이는 무한 히든마르코프 모델의 주요 구성 요소입니다. Fig. 3은 HDP 프로세스를 보여줍니다.



**HDP의 세그룹을 가진그래픽 모델,**Fig. 3

1. **잠재 의미 인덱싱(Latent Semantic Indexing, LSI)은 특이값 분해(Singular Value Decomposition, SVD)라는 수학적 기법을 사용하는 모델입니다. 행의 수를 줄이는 동시에 열**간의 유사성 구조를 보존합니다. 행은 고유한 단어를, 열은 각 문서를 나타냅니다. LSI는 분포 가설에 기반하여 작동하며, 즉 의미가 가까운 단어는같은 종류의 텍스트에서 발생할 것이라는 가정하에 동작합니다. LSI는 개별 단어 대신 통계적으로 유도된 개념적 인덱스를 검색에 사용함으로써 어휘 일치와 관련된문제를 극복합니다. LSI는 함께 사용되거나 같은 맥락에서 사용되는 단어는 유사한 의미를 가진다는 원칙에 기반합니다. 달리 말해, 동의어는 자주 함께 사용된다는 가정하에 동작합니다.

*관련문헌 고찰*

이 섹션에서는 트위터 데이터를 활용한 관련 연구를 다수 주제 모델링 기법을 통해 검토하고, 주제 모델링 기법에 대한 종합적인 검토로 마무리됩니다.

1. **트위터 데이터 연구**

트위터와 같은 마이크로블로깅 웹사이트는 특히 감정 분석 및 의견 마이닝 분야에서 많은 데이터마이닝 연구의 중심이 되어 왔습니다. 다양한 기술과 도구가 적용되어 이 대규모 데이터 세트를분석하여 의미 있는 트렌드를 발견하고, 연구자들에게 관심 있는 분야에서 예측 및 전망을 수행하기위해 노력해 왔습니다. 이러한 기술에는 머신러닝과 통계 방법 등이 포함되지만 이에 국한되지 않습니다.연구자들은 이러한 데이터를 효과적으로 관리하기 위해 다양한 데이터 마이닝 도구를 탐구하고 있습니다. 연구자들은트위터 데이터를 연구의 입력 원천으로 사용하여 다양한 주제를 다루어 왔습니다. 연구자들이 과거 몇년간 집중해 온 분야 중 일부는 질병 감지, 예측 [8-10] 및 모니터링 [11],[12]을 위한 감정 분석, 선거 예측 [13-18] 등이 있습니다. 또 다른 주목할 만한분야는 [19-21]의 연구에서 볼 수 있는 주식 시장 예측입니다. 다른 주목할 만한 분야로는 범죄 예측 [22]과 스팸 감지 [23]가 있습니다.

1. **질병 탐지: 뉴욕 시의 인플루엔자 사례 수준을 예측하기 위해 트위터 데이터를 사용하는 것을 평가하기 위해, [8]의 연구자들은 선형 회귀,**포아송 모델 및 벡터 맵과 같은 기법을 제안된 지역의 사용자로부터 수집된 트위터 데이터셋에 적용했습니다. 인플루엔자 사례를 모니터링하기 위해, [11]은지리 정보 시스템(GIS)을 사용하여 데이터를 타겟팅, 필터링 및 정규화했습니다. 또한, 그들은 지원 벡터 머신(SVM)을 분류기로 사용하는 포괄적인 데이터 마이닝프로세스를 적용했습니다. 한편, [12]의 저자들은 트위터 데이터를 인플루엔자 사례의 예측 및 모니터링에서 암 환자와 그들이 사용하는 치료법의 모니터링으로 연구를확장했습니다. [9]의 연구자들은 랜덤 포레스트 분류기를 사용하여 트위터 사용자의 우울증 및 외상 후 스트레스 장애와 같은 정신 질환의 발병을예측할 수 있었습니다. Python과 R을 분석에 사용했습니다. [10]은 질병 발발의 조기 탐지를 위해 계층적 클러스터링을 제안했습니다. 이는 이러한 사건을 효과적으로 관리하기 위한 것입니다.
2. **선거 예측: [13]은 이탈리아 사용자의 트윗을 분석하기 위해**상대적 지원 매개변수를 가진 통계 도구를 사용했습니다. 그들은정당 지도자의 트윗을 최종 선거 결과와 비교했습니다. 트윗의양과 속도가 선거 결과에 huge한 영향을 미친다는 것을 발견했습니다.

2012년 미국 대통령 선거와 2013년 카르나타카 주 의회 선거의 결과를 예측하기 위해, [14]는 주성분분석(PCA)과 서포트 벡터 머신(SVM)을 결합하여 분석에 활용했습니다. 이 연구에서 사용된 다른 기술로는 Naive Bayes,Maximum Entropy, 그리고 인공 신경망이 있습니다. 실험 결과, 서포트 벡터 머신은 조사된 다른 기술들보다선거 결과 예측에 우수함을 입증했습니다. 이전 보고서와 달리, [15]는 2010년 미국 의회 선거와 관련된트위터 데이터 분석 결과와 선거 결과 사이에 상관관계가 없음을 발견했습니다. [16]에서는 2012년 미국 대통령선거에 대한 공중의 의견을 실시간으로 분석하기 위해 트윗 분류를 위한 Naive Bayes를 제안했습니다. [17]과[18]에서 수행된 연구는 각각 2017년 프랑스 대통령 선거와 2011년 싱가포르 대통령 선거를 예측하기 위해감성 분석을 활용했습니다. 전자는 후보자의 인기를 예측하기 위해 마이크로블로그 트위터에서 데이터를 추출했으며, 후자는 후보자별 득표율을 예측하기 위해 가중치 기술을 사용했습니다.

1. **주식 시장 예측: 스쿠자와 로마노프스키[19]는 트위터 데이터를 분석하여 주식 시장을 예측하기**위한 시스템을 설계하고 구현했습니다. 3개월 동안 수집된 트윗을 분류하기 위해 머신러닝기법인 나이브 베이즈가 적용되었습니다. 또한, [20]은 트위터 데이터셋을 외부 데이터로 하는선형 회귀 모델을 적용하여 주식 시장 지표를 예측했습니다. 그들의 초기 결과는일일 트윗 수와 특정 주식 시장 지표 간에 각 수준에서 상관관계가있음을 보여주었습니다. [21]의 저자들은 트위터에서 추출한 트윗에 대한 감성 분석 및감독 학습 원칙을 적용하고, 회사의 주식 시장 움직임과 트윗의 감성 간상관관계를 분석했습니다. 그들의 결과는 주식 가격과 사용자의 트윗 간에 상관관계가 있음을 보여주었습니다.
2. **주제 모델링 연구**

Topic modelling finds its application in many diverse fields of life endeavour. One of such is in news, and poetry like the Urdu text modelling [24] in which the LDA, LSI and HDP models were applied and compared. Findingsshowed that LDA did better than LSI, and both did well with the news corpus text, but none of the models excelled in the raw poetry data. Another area of topic modelling application is in legal document, such asin the case of the Latvian legal document [25]. LDA, LSI, and HDP were also employed in the modelling and compared. Though, they all yielded negative coherence values, which was used as a measure of goodness, the LDA slightlyout-performed the other two. Legal document modelling in the British legislation [26], was also modelled using LDA, HDP, and Saffron, a domain modelling software. Saffron excelled over LDA, and HDP in that order. Topic modelling can also be usedto extract topics from a multilingual survey corpus, and automatically classify them [27]. LDA and HDP were used, and HDP was able to automatically group a good number of topics. Topic modelling using LDA is also useful in image and music domain [28].

1. *문서 모델링 기술 리뷰*

마이크로블로깅 데이터인 트위터의 경우, 문서 간 구조를 공유하는 계층적 모델인 LDA와 같은모델이 필요합니다. 이 모델은 관심 대상이 구조적 표현을 보이는 경우에 적합합니다. 코퍼스내의 주제 빈도는 파워 로 분포를 따릅니다. 즉, 몇 개의 주제가 더자주 발생하여 높은 확률을 가지지만, 대부분의 주제는 낮은 확률을 가집니다. 이는 불균형을초래합니다. 제한 볼츠만 머신(RBM)의 수정 버전인 다양화된 RBM(DRBM)은 이 문제를 해결하기 위해숨겨진 단위를 다양화하여 주요 주제뿐만 아니라 덜 중요한 주제도 표현하도록 합니다. LDA는확률적 디리클레-다항 분포 설정을 사용하지만, 파워 로 분포, 즉 언어학에서 잘 알려진지프의 법칙을 포착하지 못합니다. 이는 빈번한 결과가 확률 질량을 흡수하는 확률 과정인"부자가 더 부자가 되는 과정"과 같은 확률 과정에 의해 생성됩니다. 따라서 저자들은계층적 피트만-요르(PY) 과정을 기반으로 한 새로운 확률적 주제 모델링 접근 방식을 제안했습니다.LDA와 PY는 문서에서 주제가 생성되는 방식에서 차이가 있습니다. LDA는 문서의 각 단어에서주제를 생성하지만, PY에서는 중국 레스토랑 프로세스(CRP)라는 테이블에서 주제가 생성됩니다. CRP의 변형인 중첩중국 레스토랑 프로세스(nCRP)는 계층적 디리클레 프로세스(HDP)의 파생물입니다. 아메드, 홍, & 스몰라[31]는 중국레스토랑 프랜차이즈(CRF) 모델을 제안하여 트위터 마이크로블로그와 사용자 위치의 공동 분포를 모델링하고 위치불확실성을 40% 감소시켰습니다. CRP와 CRF는 문서가 단일 주제 혼합에서 생성될 수 있도록허용하지만, 주제 간 관계는 제공하지 않습니다. nCRP는 트리 구조를 제공하여 부모 주제가자식 주제보다 더 일반적이라는 의미론을 제공합니다. 문서 클러스터링과 주제 모델링은 tf-idf 메트릭과단어 임베딩 모델과 같은 기술의 조합을 사용하여 달성할 수도 있습니다. 온라인 주제모델링은 온라인 소스를 통해 텍스트 스트림과 그 변화가 수신되면 자동으로 LDA를 사용하여클러스터링됩니다. 새로운 데이터[32]를 사용하여 새로운 생성 과정을 학습하는 데 사용되는 비마르코프 온라인LDA 기브스 샘플러 주제 모델(OLDA)이 현재 모델과 LDA 결과와 동일한 데이터셋과 비교되었습니다.

[33]의 저자들은 트위터와 레딧의 데이터셋을 사용하여 문서와 주제 모델링을위한 워드 임베딩, doc2vec, tf-idf 등의 클러스터링 기법을 평가하고LDA와 비교했습니다. 그들은 워드 임베딩이 문서 클러스터링의 기반으로 효과적으로사용될 수 있으며, doc2vec과 tf-idf 가중치 평균 워드 임베딩표현이 문서 클러스터링 작업에서 단어 임베딩 벡터의 평균보다 더나은 결과를 제공한다는 것을 관찰했습니다. tf-idf 방법과 LDA는 500자이상의 데이터 크기에서 word2vec과 비교 가능한 결과를 제공했습니다. 자동문서 요약은 주제 모델링의 또 다른 응용 분야이며, 이는인간 수동 요약과 비교할 때 평가하기 어렵습니다. [34]에서는 ROUGE(Recall-OrientedUnderstudy for Gisting Evaluation)를 사용하여 인간 주석이 붙은 요약과LSA, LDA, Word2Vec, Doc2Vec의 네 가지 의미 모델 및tf-idf의 하나인 빈도 기반 모델을 문서 특징 추출에 사용한요약과 비교했습니다. LDA 주제 모델링의 또 다른 변형은 LDA의불투명한 주제 유형의 매개변수화 대신 임베딩 공간의 다변량 가우시안분포의 범주형 분포로 대체하는 것입니다[35]. 여기서 문서는 단어 유형의시퀀스가 아닌 단어 임베딩 시퀀스로 구성된다고 가정합니다. 가우시안 LDA의단어들은 사후 예측 분포가 ihnen 할당한 밀도에 따라 순위가 매겨집니다.

딥 뉴럴 네트워크(DNN)는 이미지 처리, 음성 및 얼굴 인식 분야에서 성공적으로 적용되어 왔습니다. 신경망 기반 시스템은 자연어 처리(NLP)에서도 효율적임을 입증했습니다. DNN은 단어 시퀀스를 과거 단어의 역사에서 예측할 수 있는 능력으로 단어 수준의 모델링에도 적용되었습니다. 계층적 순환 신경망(HRNN) 모델[36]은 문장 수준과 단어 수준의 언어 모델을 근사화하기 위한 두 단계 학습 접근 방식을 제공하며, 파이프라인 스타일로 수렴합니다.

문서에서 소셜 미디어의 주제를 모델링하는 것은 온라인 사용자들 간의 논의 주제의 유사성을 이해하기 위해 인기를 얻고 있는 추세입니다. 동일한 유형의주제를 가진 저자들 간의 커뮤니티 관계를 결정하고, Topic-Link LDA 알고리즘을 사용하여 이를 통합하는 것은 [37]에서 시도된 것입니다. 이는 주제뿐만 아니라저자들 간의 관계를 이해하려는 시도입니다. 전통적인 LDA 주제 모델링과 달리, 정보 필터링 모델인 최대 일치 패턴 기반 주제 모델은 문서에서단일 주제가 아닌 여러 주제에 대한 사용자의 관심을 모델링하려는 것입니다 [38]. 최대 일치 패턴은 문서를 효율적으로 표현하고 순위를 매기기 위해 사용됩니다.

재료 및 방법

데이터의 사전 처리 과정에서 문서가 서로 의미상 일관성을 가진 다양한 주제로 구성되어 있음을 관찰해야 합니다. 주제 모델링의 표준 인간평가 방법은 주제들 간의 의미적 일관성을 평가하기 위해 설계되었습니다. 다음 섹션에서 우리는 문서의 코퍼스를 여러 주제 모델링 기법에 적용하여관찰된 주제들을 발견하고, 주제들 간의 일관성을 결정합니다. Twitter와 Kaggle 머신러닝 저장소에서 가져온 문서들은 다양한 주제 모델링 알고리즘을 사용하여 학습 및 평가됩니다.

1. *주제모델링 방법론*

세 가지 주제 모델링 기법이 사용되고 비교되었습니다. LDA, LSA/LSI, HDP가 바로 그것입니다. 이 세 가지 기법을 사용한 이유는perplexity와 coherency를 평가 기준으로 삼아 어떤 기법이 가장 좋은 결과를 내고 주제를 더 잘 모델링하는지 알아보기 위함입니다. 이모델들은 비지도 학습 접근법을 통해 개발되었습니다. 또한, 데이터 임베딩 기법 두 가지도 사용되고 비교되어 데이터 임베딩 기법이 주제모델에 미치는 영향이 유의미한지 확인했습니다. 사용된 데이터 임베딩 기법은 bag of words (BOW)와 term-frequency inverse-document-frequency (tf-idf) 지표, 워드 임베딩 모델입니다.

Gensim 및 Mallet 프레임워크는Python용 Jupyter 노트북에서 Python3의 프로그램 개발 환경으로사용되었습니다. R Studio는 실시간트윗을 검색하는 데 사용되었습니다.다음 섹션에서는 이러한 실험에 대한 설명이 제공됩니다.

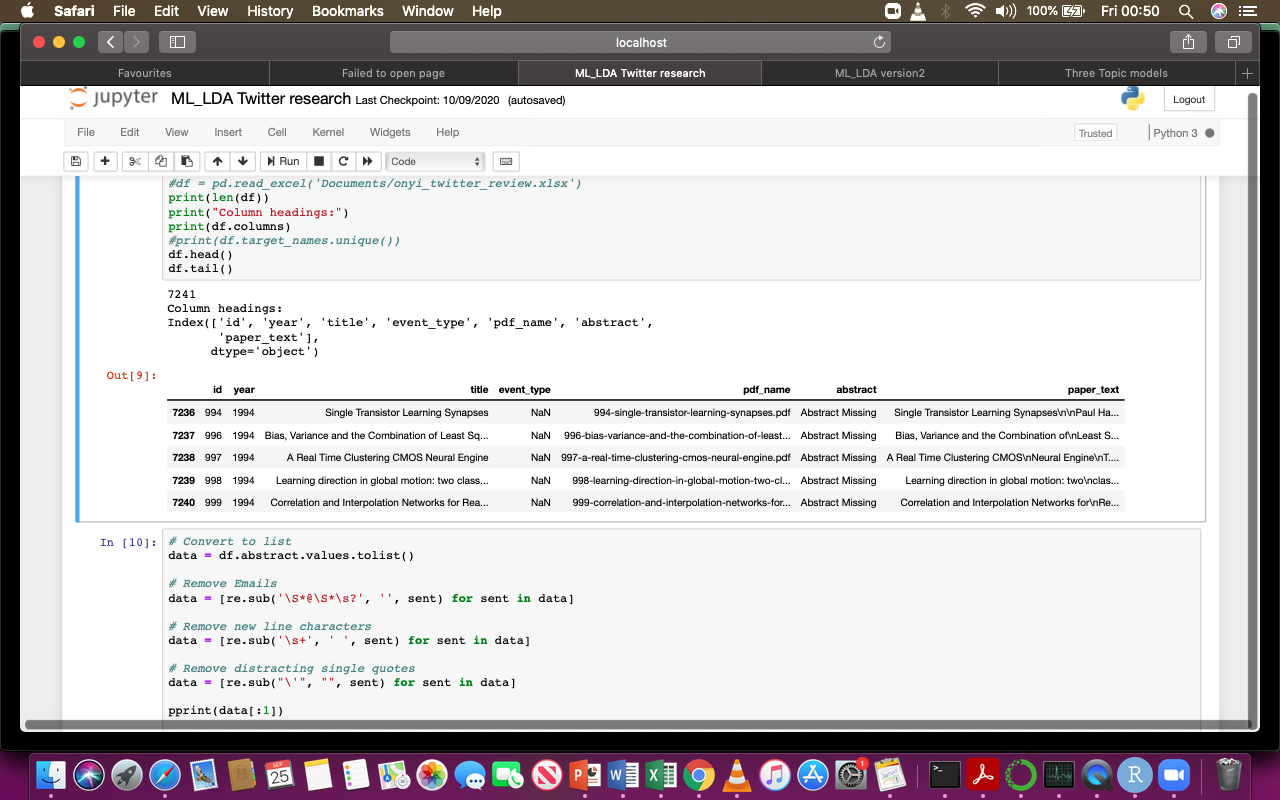
1. **데이터 소스**

연구에 사용된 두 가지 주요 데이터 원천은 Kaggle과 사용자 트윗이었습니다. 실험은 1975년부터 2017년까지의 Neural Information Processing Systems(NIPS) 회의 논문을사용해 진행되었습니다. 총 7,241개의 논문이 Kaggle 머신러닝 데이터셋에서 다운로드되었습니다(도표 4 참조). NIPS는 딥러닝, 컴퓨터 비전, 인지과학, 강화학습 등 다양한주제를 다루는 세계적인 머신러닝 컨퍼런스 중 하나입니다[39]. 이 데이터셋에는 1975년부터 2017년까지의 모든 NIPS 논문의 제목, 저자, 초록이 포함되어 있으며,총 7,241개의 문서로 구성되어 있습니다. 다른 데이터셋은 머신러닝과 AI에 대한 사용자 트윗입니다. 이 데이터셋에는 R 프로그래밍에서 Twitter API 키를사용해 트위터에서 수집한 20개 주요 AI 및 머신러닝 관련 계정의 35,860개의 비정형 트윗 데이터가 포함되어 있습니다(표 1 및 도표 5 참조).

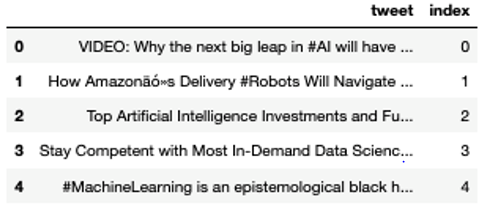
표 I

트위터 계정 정보

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SN** | **핸들** | **소유자 이름** | **팔로워** | **안녕하세요, 어떻게 지내시나요? 잘 지내시길 바랍니다.** | **rf** | **특별 관심** | **트윗 총 수** | **트윗 검색 건수** |
| 일 | @ylecun | **이안르쿤** | 238.5k | 380 |  | 광학 문자 인식 및합성곱 신경망(CNN)을 이용한 컴퓨터 비전 | 육백칠십 | 652 |
| 이 | @analyticsindiakr | 인도 분석 잡지 | 12.2k | 353 |  | 인도 AI/ML 열성가 모임 | 17.9k | 1,798 |
| 삼 | @앤드류영 | 응 앤드류 | 552.7천 | 540 |  | 자연어 처리, 심층 학습 | 1,324 | 447 |
| 4 | @karpathy | 안드레이카파시 | 232.7k | 605 |  | 테슬라의 신경망 자율주행 전문가 | 6687 | 3959 |
| 5 | @딥러닝007님 | 임티아즈아담 | 33.6k | 118.3k |  | 인공지능 기술에 대한 통찰 | 125.9천 | 7,355 |
| 6 | @sallyeaves | 에이브스 교수님 | 115.3k | 99.2k |  | 분산원장, 인공지능,클라우드 컴퓨팅 | 114.4천 | 만 |
| 7 | @kirkdborne | 박사 키르크 본 | 273k | 9,510 |  | 빅데이터, 데이터 과학, 인공지능. | 129.9천 | 8,863 |
| 8 | @파울라픽카드 | 파울라 피카르 | 60.4천 | 아홉천팔십구 |  | 기술 분야의 여성, 디지털 전환, 인공지능, 사이버 보안 | 34.9천 | 만 |
| 9 | 반룬로널드 | 반론 | 237.6천 | 180.7k |  | 인공지능,빅데이터,사물인터넷, 머신러닝, 분석. | 93.5천 | 만 |
| 10. | 스피로스마르가리스 | 스피로스마르가리스 | 106.2천 | 15.7k |  | 금융기술,인공지능(AI), 블록체인 | 208.8k | 15,000 |
| 11 | @drfeifei | 피피교수님 | 1,519 | 469 |  | 인공지능 연구를 통한 인간 조건의 향상 | 374.4k | 516 |
| 12 | @tamaramccleary | 타마라맥클레어 | 306.8천 | 207.1천 |  | 소셜 미디어 분석, 소셜 미디어 전략 및 인플루언서 프로그램. | 108.6천 | 삼천육백육 |
| 13 | @EvanKirstel | 이반키르스텔 | 296.4k | 273.2만 |  | 사회적 미디어마케팅, 사물인터넷,클라우드, 통신, 5G | 988.7k | 11,478 |
| 14 | 마이크쿠다지 | 마이크콰다지 | 156.9k | 3,134 |  | 인공지능 및 로봇 공학의 발전 | 54.7천 | 11,169 |
| 15 | @hmason | 힐러리 메이슨 | 126.7천 | 1,870 |  | 로봇 및 인공지능 연구, 데이터 사이언스 인 레지던스 | 19.4k | 64 |
| 16 | @antgrasso | 안토니오그라소 | 179k | 44.9k |  | 디지털 전환, 모바일 DevOps, 디지털 토큰 및 AI 마케팅 | 41.7k | 15,000 |
| 17 | 피셔85님 | 마이클 피셔 | 89k | 11.3k |  | 사이버 보안 및 인공지능 혁신 | 32.4k | 5,697 |
| 18 | @앤디피츠 | 인공지능앤디 피체 | 90.7k | 41.2천 |  | 2020년 인공지능 발전 전망 | 4,790 | 331 |
| 19 | @jblefevre60 | 장-바티스트레페브르 | 80.2천 | 5,551 |  | 인공지능, 머신러닝, 클라우드 분야의 혁신 | 183.9 | 15,000 |
| 이십 | @PierrePinna | 피에르 피나 | 53.1천 | 50.1k |  | 로봇 공학 분야의 혁신과 신기술 | 79k | 4,989 |
|  |  |  |  |  |  |  | **총합** | **135,860** |



**Fig. 4.구조화된 문서**(NIPS 논문)



**Fig.5.** 비정형 문서 (트윗)

1. **절차**

파이썬의 Gensim 모델은 트위터 원본 데이터셋과NIPS 회의 논문에서 주제 모델링 실험에사용됩니다. 두 실험 모두 다음 절차로 설명됩니다:

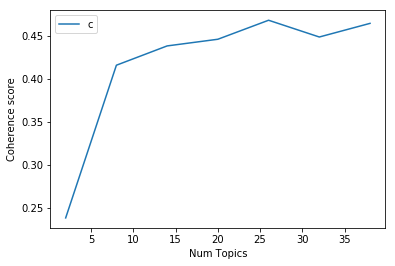
1. **데이터 로딩: 모든 데이터 파일은 쉼표로 구분된 값(csv)으로 저장되었으며, Pandas 라이브러리를 사용하여 프로그램에 읽혀졌습니다.**Fig 4와 Fig. 5는 각각의 실험에서 사용된 두 데이터셋의 Pandas가 읽은 머리 부분(Head section)을 보여줍니다.
2. **데이터 정제: 자연어 처리 도구 키트(NLTK)에서 제공하는 불용어 제거기를 사용하여 문서에**특정 의미를 더하지 않는 단어를 제거합니다. 텍스트 사전 처리에는 Spacy의 en(영어)모델이 사용되며, 이는 표제어 추출을 포함합니다. 표제어 추출은 단어를 그 기본형으로변환하는 데 도움이 됩니다. 예를 들어, 'walking'은 'walk'로, 'machines'는 'machine'으로 표제어 추출됩니다.
3. **구문 모델링: 이중어와 삼중어: 문서에서**일부 단어는 두 단어 또는세 단어로 자주 함께 나타납니다.이러한 단어를 인식하는 것이 중요합니다.Gensim의 구문 모델은 이중어와 삼중어를 구축하는 데 사용됩니다.
4. **주제 모델링을 위한 사전 및 코퍼스**생성. LDA 주제 모델의 두 가지주요 입력은 사전과 코퍼스로, 이를 위해사전 및 코퍼스를 생성했습니다. Gensim은 문서내 각 단어에 대해 고유한 ID를 생성합니다.

기타 절차에는 기본 모델 성능, 하이퍼파라미터튜닝, 최종 모델 및 결과 시각화가 포함됩니다.

1. *최적의 주제 수를 어떻게 결정하나요?*

26개의 주제로 실험을 반복하여 더 잘 알고 있습니다. 이 함수는 여러 모델을 학습시키고 각각의 일관성 값을 출력합니다. 곡선이 평평해지기 시작하기 전에 일관성 값이 가장 높은 주제는 일반적으로 선택하는 것이 가장 좋습니다. 이 경우, 그림 6에 표시된 바와 같이 26번 주제를 선택했습니다. 이 주제는 인덱스 위치 4를 가지고 있습니다.

파이썬 프로그램은 LDA, LSI 및 HDP 모델에 대해 주제 수 2, 8, 14, 20, 26, 32, 38, 44 및50에 걸쳐 반복적으로 실행되었으며, 두 데이터 세트를 순차적으로 사용하고 단어 임베딩을 교체했습니다. 결과는 이후 섹션에서 보고됩니다. 다양한 주제 수를사용하면 단일 지점을 기준으로 판단하는 대신 주제 모델의 성능에 대한 진정한 그림을 얻을 수 있습니다. 결과는 표와 시각화를 통해 제시됩니다.



**Fig. 6.주제 수를**결정하기 위한 플롯

1. *주제 모델링 평가 지표*

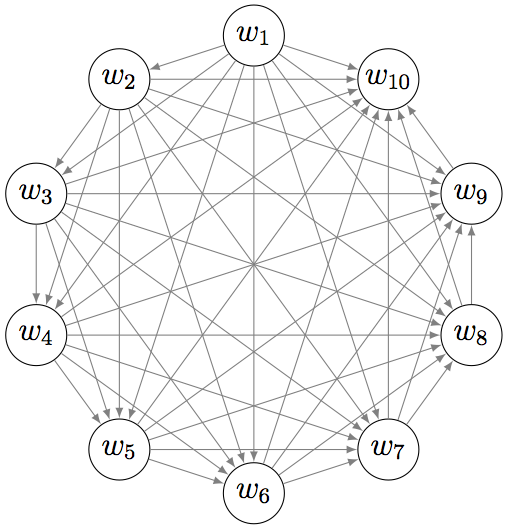
LDA 주제 모델링은 다른 비지도 학습 알고리즘과 마찬가지로평가하기 어렵습니다. 이는 이를 평가할 수 있는 명확한기준이 없기 때문입니다. 주제 모델링 성능을 평가하는 다양한방법이 있습니다. 이는 일관성 측정 [40-42], 혼동도 [43],Chib-style 추정기 및 "왼쪽에서 오른쪽으로" 알고리즘 [44], 시각화방법 등을 포함합니다. 주제 모델의 성능을 평가하는 두가지 인기 있는 방법은 일관성과 혼동도입니다. 일관성 측정은주어진 주제에서 상위 단어 쌍 간의 유사도의 평균으로,주제의 품질을 측정하는 데 채택되었습니다 [40]. 이는 주제모델이 결과의 해석 가능성을 보장하지 않기 때문입니다. 일관성측정에는 내재적 및 외재적 두 가지 유형이 있습니다[41]. 두 가지 모두 주제를 설명하는 데 사용되는단어 w1,..,wn에 대한 쌍별 점수의 합을 계산합니다(도표 7참조). 이는 주제 p(w|k)p(w|k)에서 빈도순으로 상위 n개의 단어를설명하는 방정식 2와 같습니다. 이 측정치는 도표 7의 모든 엣지의 합으로 간주됩니다.

Coherence = (2)

1. **Intrinsic (Umass)**: this measure compares a word only to the previous and next words, respectively. The words have to be an ordered set for this to work. It uses a pairwise score function, an empirical conditional log-probability with smoothing count by adding one to D(wi,wj) in order to avoid calculating the log of zero. Where p(w) is the probability of seeing word, wi in a random document, and p(wi,wj) is the probability of both words occurring in a random document.

The Umass score is given as ScoreUMASS(wi,wj)= (3)

### **Extrinsic Measure (UCI)**: uses a pointwise score function known as Pointwise Mutual Information (PIM) in which single words are paired with each other. The UCI score is given in (4). ScoreUCI(wi,wj)= (4)



**Fig.7.**주제내단어 연결 (출처: Pleple [41])

주제 일관성 측정은 인간 해석 가능성에 기반하여 다양한 주제 모델을 비교하는 좋은 방법입니다.

1. **LDA의 perplexity 점수는 보류된**테스트 데이터의 로그-우도(log-likelihood)가 가장일반적인 평가 방법입니다 [45].지도 학습과 마찬가지로, LDA의출력 또는 학습 데이터셋은각각 훈련 세트와 테스트세트로 분할할 수 있습니다.LDA에서 테스트 세트는 관찰되지않은 문서, wd의 모음이며,q는 모델 주제 행렬을설명하고, 하이퍼파라미터 a는 문서의주제 분포를 나타냅니다. 로그-우도는(4)에서와 같이 평가됩니다. 더 높은 로그-우도는 더 나은 모델을 의미합니다.

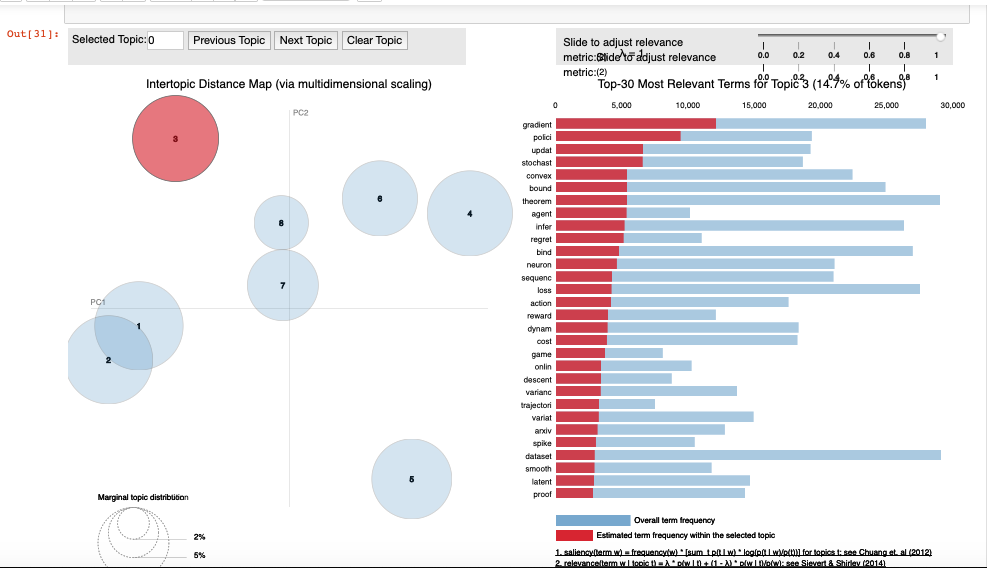
L(w) = logp(w|q,a)= (4)

LDA 모델은 주제 모델에서 보류된문서 Wd에 대해 전통적으로 사용되는퍼플렉서티 방법을 사용합니다. 퍼플렉서티는 모델이새로운 텍스트 문서를 일반화하고 예측하는능력을 측정하는 지표입니다 [43]. 이는문서의 임의의 위치에서 발생할 수있는 동등한 가능성이 있는 단어의수를 나타냅니다. LDA 모델은 입력으로단어 집합을 받은 후 퍼플렉서티 값을 반환하는 log\_perplexity 방법을 사용합니다.

Perplexity(test\_set w) = (5)

혼란도값이낮을수록모델이더 좋습니다 [43,45].

1. **시각적 평가: LDA 시각화 도구인 pyLDAvis는 토픽 모델의 결과를 평가하는 또 다른 방법입니다. 이는 LDA**모델에서 추출된 토픽을 원형 또는 버블로 표현하는 Python 상호작용 라이브러리입니다. 원이 클수록 해당 토픽을 구성하는단어의 비율이 높습니다. 버블의 수는 LDA 모델이 생성해야 할 토픽의 수에 따라 결정됩니다(참조 그림 8).버블은 토픽의 수에 따라 순차적으로 번호가 매겨집니다. 한 버블에서 다른 버블로 이동하면 토픽이 동적으로 변경되며,각 토픽을 구성하는 단어와 코퍼스 내 각 단어의 전체 빈도(파란색 막대 표시), 특정 토픽에 의해생성된 용어의 추정 횟수(빨간색 막대 표시)가 표시됩니다. 예를 들어, 'gradient'라는 단어는 약 28,000개의 위치에 등장했지만,토픽 3에서는 약 12,000번 사용되었습니다. 버블이 서로 멀리 떨어져 있을수록 토픽의 유사성이 낮고, 가까울수록 유사성이높습니다. 각 토픽의 용어를 통해 해당 토픽이 무엇인지 판단할 수 있습니다. 기본적으로 pyLDAvis는 토픽 내 상위 30개 가장 관련성 높은 용어를 표시합니다.



**Fig.8.**pyLDAvis를이용한토픽 모델의 시각화

IV. 결과 및 논의

이 섹션에서는 주제 모델의 평가 방법과 다음실험의 결과를 제시합니다; (i) LDA, LSI 및HDP의 세 가지 주제 모델링 기법을 비교하여가장 우수한 알고리즘을 결정합니다. (ii) 주제 모델링을위한 더 나은 임베딩 기법으로 Bag ofWords (BOW) 임베딩과 Term Frequency Inverse DocumentFrequency (TF-IDF)를 결정합니다. (iii) 이 섹션에서는 또한 Gensim과 Mallet 모델의 비교 결과를 제시합니다.

1. *실험 결과*
2. **비정형 데이터 vs 정형 데이터**

Table 2는 두 데이터셋인 비정형 트윗과 구조화된 NIPS 데이터셋에 대한 LDA 평가 결과의 요약을 보여줍니다.

표 II

두 데이터셋의 평가 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 데이터셋 | 문서 수량 | 일관성UMass 점수 | 일관성UCI점수(c\_v) | 혼란 |
| 실시간 트윗 | 135,924 | -13.53022 | 0.3945671 | -22.73715 |
| NIPS 논문 | 7,241 | -0.72836806 | 0.457887571 | -8.688008 |

NIPS 데이터셋은 일관성에서 더 나은 성능을 보였습니다. 이는 데이터셋이 구조화되어 있어 알고리즘이 주제를 더 잘 조직하기 쉽기 때문입니다. 그러나 원시트윗은 혼돈도(perplexity) 성능에서 더 나은 결과를 보였습니다. 다만, 일부 저자들은 혼돈도가 성능을 측정하는 좋은 지표는 아니라고 주장하기도 합니다. NIPS 데이터셋은UCI 일관성 점수 기준으로 원시 트윗보다 86% 더 나은 성능을 보였으며, U\_mass 일관성 점수 기준으로 5.4% 더 나은 성능을 보였습니다.

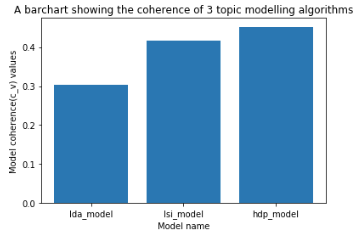
1. **LDA, LSI,HDP에 대한트윗의 BOW 주제 결과**

표 III

세 가지 모델의 두 데이터셋에서의 성능 비교

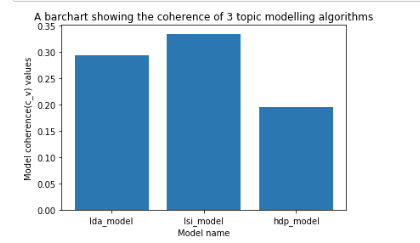
|  |  |  |
| --- | --- | --- |
| **모델** | **NIPS 일관성** | **트윗의 일관성** |
| LDA | 혼란도: -7.570  0.3026072675 | 혼동도: -7.0080  0.29372262845 |
| LSI | 0.4166050449 | 0.33385338957 |
| HDP | 0.45218042459 | 0.195466477917 |

모든 모델은 NIPS 데이터셋에서 비구조화된 트윗보다 더 잘 수행되었습니다. 이는 표 2의일반적인 결과와 일치합니다. 일관성과 LDA 퍼플렉서티는 이를 입증합니다. HDP 모델은 NIPS 데이터셋을사용하여 LSI 및 LDA 모델보다 fig. 9에 표시된 대로 순서대로 더 잘 수행되었습니다.



**Fig. 9. LDA,LSI 및 HDP**모델의 NIPS 일관성 값

LSI는 트위터 데이터셋을 사용하여 LDA와 HDP를 차례로 이겼습니다. 자세한 내용은 그림 10을 참고하세요.



**Fig. 10.LDA, LSI**및 HDP모델의 트윗일관성 값

표 4

트윗의 세 모델과 주제별 평가 점수

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **주제 없음** | **LDA\_혼란도 지수** | **LDA\_CV** | **LSI\_CV** | **HDP\_CV** |
| **이** | -6.992287979 | 0.26038149399 | 0.2109477416 | 0.1971698105 |
| **8** | -7.008076761 | 0.29372262845 | 0.3338533896 | 0.1954664779 |
| **14** | -9.042080899 | 0.35260845743 | 0.3024220818 | 0.1964664779 |
| **이십** | -11.41537552 | 0.39173503481 | 0.3092057430 | 0.1948254684 |
| **이십육** | -12.62313976 | 0.42963160923 | 0.2963431871 | 0.2011519489 |
| **삼십이** | -13.85463406 | 0.43276902837 | 0.3158016858 | 0.1954439597 |
| **38** | -15.02677988 | 0.45899579357 | 0.3229799664 | 0.1957414457 |
| **44** | -16.25851165 | 0.46683228728 | 0.3325065428 | 0.1983655346 |
| **50** | -17.91880002 | 0.46980153149 | 0.3405411231 | 0.1965291292 |

표 4와 그림 11은 LDA, LSI 및 HDP 모델의 평가 성능 결과를 주제 수 2, 8, 14, 20, 26, 32, 38, 44 및 50에 걸쳐 보여줍니다. LDA 모델에서는 주제 수가 증가함에따라 일관성 점수가 꾸준히 증가하는 반면, 혼란도는 감소합니다. LSI와 HDP는 결과가 변동합니다. HDP는 주제 수를 결정하는 방법이 없지만, LDA와 LSI에서 주제 수를 변경할 때마다 각 프로그램 실행 시 다른 일관성 값이 생성됩니다.

**Fig. 11. LDA,LSI 및 HDP의**주제 수에 따른 일관성 점수

LDA 모델은 주제 수와 일관성 측면에서 우수하며, 모든 경우 중 하나(주제 수8인 경우)를 제외하고는 가장 높은 일관성 값을 보였습니다. 이는 4.2.1절에서 승자가 불분명한결과와 대조적입니다. LSI는 두 번째로 높은 순위를 차지했으며, HDP는 상대적으로 낮은 성능을 보였습니다.

1. *단어 빈도 수 vs 역 문서 빈도*

두 가장 자주 사용되는 코퍼스 중 더 나은 임베딩 방법을 찾기 위해, LDA Bag of Words 모델과 LDA TF-IDF 모델을 사용하여 NIPS 테스트 데이터셋 문서를 분류하는 방식으로 성능 평가가 수행되었습니다. 다음은 Table 5에서 비교된 결과입니다.

표 V

LDA-BOW 및 LDA-TF-IDF 임베딩의 결과 주제 평가

|  |  |  |
| --- | --- | --- |
|  | 혼란 | 일관성 |
| LDA-BOW | -7.739107959 | 0.49215338 |
| LDA-TFDIF | -7.811492707 | 0.25257377 |

BoW 임베딩은 일관성 값에 따르면 TF-IDF보다 50.9% 더우수한 성능을 보였으며, 또한 perplexity 평가에서도 더 나은결과를 제공했습니다. 미확인 문서를 예측할 때, BoW는 문서를주제별로 분류하는 데 0.99(99%)의 정확도를 기록하며 더 우수한성능을 보였습니다. 반면, TF-IDF는 각각 0.58(58%)의 정확도를 기록했습니다.

1. *트위터 데이터결과: Gensimvs Mallet 모델*

표 6은 Gensim 및 Mallet모델에서 각각 관찰된 Twitter 데이터셋의결과를 20주제 LDA 모델에서 비교하고있습니다. 선택된 결과는 각 주제에서0.2 또는 20% 이상의 용어가중치와 용어 확률에 기반하고 있습니다.Mallet은 13개의 용어를, Gensim은 20퍼센트 이상의 가치를 가진 9개의용어를 가지고 있습니다. Mallet과 Gensim모두 4개의 주제를 공유하고 있으며,가중치는 다릅니다. 공통된 단어는 다음과같습니다: ai, developer, machine, 그리고machinelearne. Mallet 모델은 더 많은 주제를 식별하고 더 관련성 있는 주제를 생성하는 능력에서 Gensim보다 우수합니다.

표 6

Gensim vs. Mallet 모델

|  |  |  |
| --- | --- | --- |
| 주제 | 해머 | 젠심 |
| 0 | 딥러닝 |  |
| 하루 | 0.277\*"머신러닝" |  |
| 이 | 0.224\*"머신러닝" |  |
| 3 | 0.289\*"데이터사이언스 | 0.352\*"데이터 |
| 4 | 0.321\*"https" |  |
| 5 | 0.247\*"머신러닝" |  |
| 6 |  | 0.221\*"생성" |
| 8 | 0.557\*"ai" |  |
| 9 | 개발자 0.218\* | 0.290\*"파이썬 |
| 10 | 0.339\*"기계" |  |
| 12 | 0.245\*"트위터 |  |
| 13 |  | 0.284\*"개발자"0.242\*"완료 |
| 14 |  | 0.300\*"기계"0.204\*"학습 |
| 16 | 0.259\*코드 0.214\*프로그래밍 0.206\*프로그래머 |  |
| 18 |  | 0.391\*"인공지능"0.320\*"머신러닝" |

1. *트위터가 인공지능 및 머신러닝 연구에 미치는 영향 측정*

트윗과 NIPS 논문 간의 상관관계 비교

이 섹션에서는 트윗이 인공지능(AI)과 머신러닝(ML) 분야의 연구 방향에 영향을 미칠 수 있는지 알아보려고 합니다. NIPS 논문들은 AI와 ML 분야의 연구 방향을 알려줍니다. 사람들이 트윗하는 내용이 연구자들이 연구하고 싶은 주제에 영향을 미칠 수 있을까요? 이 영향력을 측정할 수 있는 방법은 있을까요? 이 섹션은 이러한 질문에 답하기 위해 작성되었습니다.

이러한 질문에 답하기 위해, 먼저 해당주제에 대해 트윗하는 사람들을 조사하고, 소셜미디어에서의 그들의 영향력을 측정하는 방법을 찾아야합니다. 특히, 연구 커뮤니티와의 비교를 통해분석합니다. 다음으로, 트윗과 NIPS 논문 간의 공통점을 분석합니다.

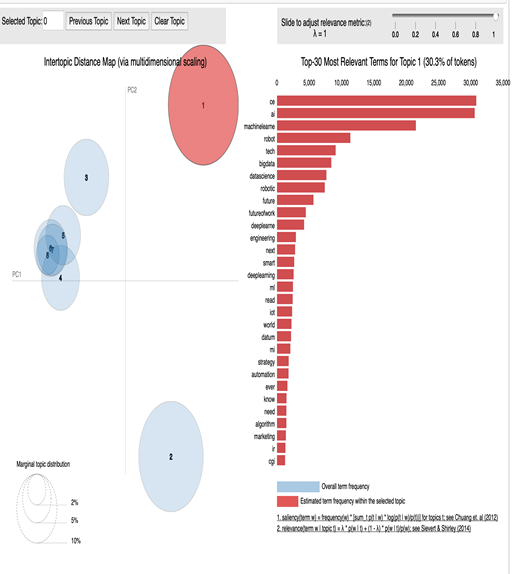
1. **트위터의 영향력과 영향자**

트윗이 AI/ML 생태계의 연구 방향에 영향을 미치나요? 만약 그렇다면, 영향력 있는 인물은 누구인가요?

Fig. 12는 8개의 주제를 가진 Twitter 주제모델의 pyLDAvis 시각화입니다. 버블(트위터 주제)에 커서를 가져가면해당 주제의 논의 내용에 대한 아이디어를 얻을수 있습니다. 예를 들어, 주제 1(토큰의 30.2%)은AI, 머신러닝, 로봇공학, 데이터 사이언스 등을 논의하고있습니다. 이는 해당 주제가 AI 및 그하위 분야에 대한 것임을 보여줍니다. 주제 2(토큰의25.6%)는 연구, 게임, 발명, 예측 등 AI분야의 혁신에 대해 논의하고 있습니다. 주제 3(토큰의12.3%)은 핀테크; 핀테크 기업, 소프트웨어 예측 모델등을 포함한 핫한 핀테크 트렌드에 대해 논의하고있습니다. 주제 4-8은 다양한程度上 겹치며, 기술 및응용 분야에서 머신러닝 기술과 스킬 개발, 건강 등 다양한 분야에 대해 논의하고 있습니다.

2020년 3분기 기준 1억 8700만 명의 활성 트위터 사용자 중, 소수만이 인플루언서, 즉 알파 그룹[46]에 속합니다. 이 그룹은 대규모 팔로워와 생성한 콘텐츠로 특징지어집니다. 이러한 콘텐츠는 주로 리트윗과 멘션을 통해 확산되며, 직접 팔로워가 아닌 다른 마이크로블로그 사용자에게도 도달할 수 있습니다. 이러한 광범위한 접촉은 인플루언서가 사용자 도달 범위에 더 큰 영향을 미치게 합니다. 일부 연구자들은 인플루언서를 의견 리더, 혁신가, 권위 있는 인물[47] 등으로 부르기도 합니다. 또한, 특정 분야의 전문가, 예를 들어 표 1에 제시된 머신러닝과 인공지능 분야의 전문가로 간주되기도 합니다.

리켈메와 곤살레스-칸테르기아니[47]는 인기, 영향력, 활동량등을 측정하기 위한 트위터 지표를제시한 표를 제공했습니다. 예를 들어,사용자의 인기를 측정할 때는 F1/F3또는 트위터 팔로워-팔로잉 비율(TFF)을 사용합니다.여기서 F1은 팔로워 수이고 F3은팔로잉 수입니다. 표 7에서는 사용자의인기를 계산하고 논의 주제에서 상위10명의 영향력을 순위매긴 후 표 1의 수정된 버전을 보여줍니다.



**트위터 주제모델의 8개**주제에 대한pyLDAvis 시각화,Fig. 12

표 VII

인기 트위터 AI 및 머신러닝 사용자

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **시리얼 번호** | **핸들** | **소유자 이름** | **TFF**(F1/F3) | **순위**(F1/(F1+F3)) |
| 하루 | @앤드류영 | 앤드루 응 | 1023.518519 | 0.999023932 |
| 이 | @ylecun | **옹레큰** | 627.6315789 | 0.998409243 |
| 삼 | @카파시 | 안드레이카파시 | 384.6280992 | 0.997406828 |
| 4 | @hmason | 힐러리 메이슨 | 67.7540107 | 0.985455394 |
| 5 | 마이크,어떻게 지내? | 마이크퀸다지 | 50.06381621 | 0.980416661 |
| 6 | @한국데이터분석 | Analytics India Magazine 인도 분석 매거진 | 34.56090652 | 0.971879232 |
| 7 | @kirkdborne | 박사 키르크 본 | 28.70662461 | 0.966337475 |
| 8 | @jblefevre60 | 장-바티스트르페브르 | 14.44784723 | 0.935266061 |
| 9 | @fisher85m | 마이클 피셔 | 7.876106195 | 0.887337986 |
| 10 | 스피로스마르가리스 | 스피로스마가리스 | 6.76433121 | 0.871205906 |

[46]의 저자들은 트윗의 영향력이 해당 트윗의 메시지에 의해 영향을 받은 사용자 수로 측정될 수 있다고 관찰했습니다.이 영향력은 강도, 인지적, 감정적, 행동적 영향으로 나타날 수 있습니다. 본 연구에서는 표 7의 상위 5인인플루언서의 H-지수, 인용 횟수 등을 고려하여 그들의 영향력을 더욱 심층적으로 분석했습니다. 결과는 표 8에 제시되어 있습니다.

표 VIII

구글 스콜라와 스코푸스에서 가장 영향력 있는 트위터인플루언서 TOP 5 (2021년 3월 1일 기준)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Google Scala** | | | **SCOPUS** | | |
| **시리얼 번호** | **트위터 계정** | **이름** | **H-지수** | **I10-INDEX** | **인용 횟수** | **H-지수** | **문서** | **인용** |
| 1. | @앤드류영 | 응 앤드루 | 130 | 279 | 166833 | 88 | 203 | 65077 |
| 2. | @ylecun | 이안르쿤 | 126 | 294 | 186756 | 74 | 192 | 72556 |
| 3. | @karpathy | 안드레이카파시 | 11 | 11 | 35093 | 12 | 13 | 18298 |
| 4. | @hmason | 힐러리 메이슨 | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제공된 지침과 언어별 노트를 준수하여 번역을 진행하겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제공해 주신 지침과 언어별 노트를 따라 정확하고 자연스러운 번역을 제공해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 지정된 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제시해 주신 지침과 언어별 주의사항을 준수하여 번역해 드리겠습니다. | 제공된 입력 텍스트가 없으므로 번역을 생성할 수 없습니다. 번역을 원하시는 텍스트를 제공해 주시면, 지정된 지침 및 언어별 노트를 모두 충족하는 번역을 제공해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제공해 주신 지침과 언어별 노트를 따라 정확하고 자연스러운 번역을 해드리겠습니다. |
| 5. | 마이크콰다지 | 마이크콰다지 | 제공된 입력 텍스트가 없으므로 번역을 생성할 수 없습니다. 번역을 원하시는 텍스트를 제공해 주시면, 지정된 지침 및 언어별 노트를 모두 충족하는 번역을 제공해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제시하신 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 지정된 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제시하신 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제공해 주신 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. | 알겠습니다. 한국어로 번역해 드릴 텍스트를 제공해 주시면, 제시하신 지침과 언어별 노트를 준수하여 번역해 드리겠습니다. |

표 8은 구글 스콜라 및 스코퍼스 웹사이트에서 제공한 상위 5명의 트위터 영향력 있는 인물들을 보여줍니다.

1. 앤드류 응은 코세라의 공동 창립자이자 스탠퍼드 대학교 컴퓨터 과학과 객원 교수입니다. 그는 스탠퍼드 머신러닝 그룹을 이끌고 있습니다.
2. Yann LeCun은 페이스북의 수석인공지능 과학자이자 뉴욕 대학교쿠란트 연구소의 실버 교수입니다.
3. 안드레이 카파시(Andrej Karpathy)는 테슬라의 인공지능 및 오토파일럿 비전부문 디렉터입니다. 그는 심층 학습 및 컴퓨터 비전에특화되어 있습니다. 그는 이전에 스탠퍼드 대학교의 연구 과학자였습니다.
4. 힐러리 메이슨은@hiddendoor 코퍼레이션의공동 창립자이자,전 Fastforward연구소의 설립자입니다.
5. 마이크 퀸다지는디지털 얼라이언스의영업 담당자입니다.
6. **구글 스칼라에서 연구 관심사에 따른 상위 5명의 머신러닝 연구자**
7. 제프리 힌튼(인용: 425,125): 그는 토론토 대학의 명예 교수입니다.
8. 요슈아 벤지오(인용: 376,447):몬트리올 대학교의컴퓨터 과학 교수
9. 로버트 티브셔라니 (인용:367,517): 스탠퍼드 대학교생물의학 데이터 과학 교수
10. 트레버 해스티 (인용: 255,606): 스탠퍼드 대학교 통계학 교수
11. 앤드루 지서먼(인용: 244,410)은 옥스퍼드 대학의 컴퓨터 과학 교수이자 컴퓨터 비전 연구자입니다.
12. **결과**

트위터에서 식별된 상위 5명의 머신러닝 영향력 있는 인물(표 8 참조)은 주로산업 기반 전문가이며, 일부는 연구 및 학계 경험과 산업 경험을 결합하고 있습니다.

구글 스칼라의 데이터와 비교할 때, 구글 프로필을 가진 상위 5명의 머신러닝 연구자는 트위터 영향력 상위 5인보다 더 많은 인용을 받은 것으로 관찰되었습니다.

대부분의 구글 스칼라에서 상위 연구자이자 머신러닝 분야의 교수 및 선두 연구자들은 트위터에서 매우 활발하지는 않지만 트위터 계정을 가지고 있습니다. 따라서 표 7의 상위 AI/ML영향력 있는 인물들은 트위터에서 활발히 활동하기 때문에 영향력을 행사할 수 있는 반면, 인용 횟수가 더 많은 학자들은 소셜 미디어에서 덜 활발하다는 것을 추론할 수 있습니다.

LDA 모델이 다른 두 모델보다 우수한 성능을 보인다는 우리의 연구 결과에 따라, NIPS 논문과 사용자 트윗 데이터를 대상으로 LDA 평가 지표를 비교해 보았습니다. 표 9는 주제 수(2-50개 주제)를 달리한 경우의 일관성과 혼란도 값을 보여줍니다. 두 데이터셋의 결과 간에 유의미한 차이가 있는지 확인하기 위해 t-검정을 적용했습니다.

표 IX

30년간의 NIPS 논문 및 트윗에서 LDA 모델과 주제에 대한 평가 점수

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 주제 수 | NIPS 논문을 위한 LDA\_CV | LDA\_CV를 이용한 AI\_ML 트윗 분석 | LDA\_혼란도  NIPS 논문들 | LDA 혼돈도 관련하여 번역을 원하시는 구체적인 트윗을 알려주세요. |
| 이 | 0.333050244307 | 0.260381 | -7.738696634 | -6.992287979 |
| 8 | 0.412244281318 | 0.293723 | -7.829689027 | -7.008076761 |
| 14 | 0.465193850187 | 0.352608 | -8.785319047 | -9.042080899 |
| 이십년 | 0.467526397358 | 0.391735 | -9.389655864 | -11.41537552 |
| 이십육 | 0.476460525678 | 0.429632 | -9.899802937 | -12.62313976 |
| 32 | 0.46242261588 | 0.4327690 | -10.4234657 | -13.85463406 |
| 서른여덟 | 0.47927547322 | 0.4589958 | -10.9348307 | -15.02677988 |
| 44 | 0.47782591896 | 0.4668323 | -11.4634772 | -16.25851165 |
| 50 | 0.46616613333 | 0.4698015 | -11.9916904 | -17.91880002 |

가설을 다음과 같이 제시하여 결과를 검증하고자 합니다:

호: NIPS 논문과 사용자 트윗에서 생성된 주제 간에는 유의미한 차이가 없습니다.

H1: 두 데이터셋의 결과 사이에 중요한 차이가 있습니다.

t-검정이 NIPS 논문 및 사용자 트윗 데이터셋에 대한 평균 차이에서 -0.01150148에서 0.11898767까지의 95% 신뢰 구간에서 t-통계량 값 1.7445와 p-값 0.09873을 반환했습니다. NIPS 논문 및 사용자 트윗 그룹의 평균은 각각 0.4489073과 0.3951642였습니다.

p-값이 0.05보다 큰 경우, 영가설(Ho)은 채택되고 대립가설(H1)은 기각됩니다. 이는 NIPS 논문과 사용자 트윗에서 얻은 결과 사이에 유의미한차이가 없음을 의미합니다. 동일한 결과는 perplexity 점수를 고려할 때도 적용되며, 이는 0.1175의 p-값을 반환했으며, 이는 통계적으로 유의미하지 않습니다.

V. 결론

이 연구는 웹에서 연구 데이터를 수집하고 통찰력 있는 데이터 분석을 수행하기 위한 도구와 방법을 연구자들이 연구 논문이 비수용되는 문제를 해결하기 위해 체계적으로 보여주기 위해 진행되었습니다. 이 연구는 40년 동안의 NIPS학술대회 논문 데이터셋과 트위터 사용자의 트윗을 사용하여 트윗이 인공지능 및 머신러닝 분야의 학술 연구 방향에 미치는 영향의 정도를 측정했습니다. 세 가지 주제 모델링 알고리즘인 LDA, LSI, HDP가 적용되었으며, 알고리즘은 다양한程度上AI/ML 주제를 성공적으로 식별했습니다. 알고리즘의 결과는 다른 주제 수에 걸쳐 비교되었으며, LDA가 가장 좋은 결과를 보였으며, 이는 문헌과 일치했습니다. 또한 LDA 모델 내에서 임베딩 방법을 비교했으며, 각각 BoW와 tf-idf였습니다. 전통적인LDA 임베딩 방법인 BoW는 코퍼스 내 더 많은 주제를 발견하는 데 더 나은 성과를 보였습니다. LDA 모델의 평가 지표를 NIPS 논문과 사용자 트윗 사이에서 비교했을 때, 조직화된 데이터셋인 NIPS 데이터셋이더 나은 성과를 보였습니다. 그러나 t-검정 결과 두 데이터셋의 결과 차이는 유의하지 않았습니다. [47]의 저자는 트윗에 기반한 LDA 측정의 한 가지 문제는 전통적인 LDA 알고리즘이 트윗보다 더 긴 텍스트에 사용된다는점이며, 특정 변형을 사용해야 한다는 점을 관찰했습니다. 이 논문의 주목할 만한 발견 중 하나는 주제 모델링을 사용하여 트윗에서 AI/ML 연구의 주요 영향자들을 식별할 수 있다는 것입니다. 트위터에서 식별된 상위 5명의머신러닝 영향자들은 주로 산업 기반 전문가들로, 일부는 연구와 학계 경험을 산업 경험과 결합하고 있습니다. 이러한 영향자들은 트위터에서 활발한 활동으로 인해 영향력을 행사할 수 있는 반면, 더 많이 인용되는 학자들은 소셜 미디어에서 덜 활발합니다.

참고문헌

B.Batrinca및P.C.Treleaven,"소셜미디어분석:기술,도구및플랫폼에대한조사",AI&Soc,[30,](https://doi.org/10.1007/s00146-014-0549-4)89–116, 2015, https://doi.org/10.1007/s00146-014-0549-4.

1. L.Bornmann과R.Haunschild,“Tfactor:트위터영향력측정지표,”MalaysianJournal of Library & Information Science, Vol. 21, no. 2, 13-20, 2016.
2. J.Ortega,"Twitter에학술지의존재와그확산(tweets)과연구영향력(citations)의관계," Aslib Proceedings, 69, 10.1108/AJIM-02-2017-0055, 2017.
3. S.하우슈타인,"학술트위터지표,양적과학및기술 연구 핸드북," 2018.
4. T. 호프만, 확률적 잠재의미 색인, SIGIR'99: 정보검색 연구 및 개발에관한 제22회 국제 ACMSIGIR 회의 논문집, pp.50–57, ACM Press, 1999.
5. D.M.Blei,A.Y.Ng및M.I.Jordan,"LatentDirichletAllocation,"JournalofMachineLearningResearch,3, 993–1022, 2003.
6. B.그륀과K.호르닉,"Topicmodels:R패키지를이용한토픽모델적합,"JournalofStatisticalSoftware,40(13), pp. 1-30, ISSN 1548-7660, 2011.
7. R.Nagar,Q.Yuan,C.C.Freifeld,M.Santillana,A.Nojima,R.Chunara및J.S.Brownstein,“2012-2013년뉴욕시인플루엔자시즌에대한일일지리코드Twitter데이터의시간적및공간적관점에서본사례 연구,” Journal of medical Internet research, 16(10), e236, 2014.
8. A.G.Reece,A.J.Reagan,K.L.Lix,P.S.Dodds,C.M.Danforth및E.J.Langer,“트위터데이터를 이용한 정신 질환의 발병 및 경과 예측,” 과학 보고, 7(1), 1-11, 2017.
9. A.아쇼크,M.구루프라사드,C.O.프라카시및S.S.실라자,"트위터데이터를이용한질병감시및시각화에대한머신러닝접근법",2019 국제 계산 지능 데이터 과학 회의(ICCIDS), IEEE, pp. 1-6, 2019.
10. C.Allen,M.H.Tsou,A.Aslam,A.Nagel,J.M.Gawron,"트위터데이터를이용한인플루엔자다중척도감시를위한GIS및 머신러닝 방법 적용", PloS one, 11(7), e0157734, 2016.
11. K.Lee,A.Agrawal및A.Choudhary,“트위터데이터를이용한실시간질병감시:인플루엔자및암에대한 시범 연구,” ACM SIGKDD 국제 지식 발견 및 데이터 마이닝 회의 회의록, pp. 1474-1477, 2013.
12. G.Caldarelli,A.Chessa,F.Pammolli,G.Pompa,M.Puliga,M.Riccaboni및G.Riotta,"트위터데이터를이용한이탈리아정치선거의다층적지리적연구",PloSone,9(5),e95809,2014.
13. M.Anjaria및R.M.R.Guddeti,"감독학습을이용한트위터데이터의영향요인기반의견분석,"2014제6회국제 통신 시스템 및 네트워크 회의(COMSNETS), IEEE, pp. 1-8, 2014.
14. D.Gayo-Avello,P.T.Metaxas및E.Mustafaraj,"트위터를이용한선거예측의한계",AAAI웹로그및소셜 미디어 국제 회의 제5회 회의록. 인공지능 진흥 협회, 2011.
15. H.Wang,D.Can,A.Kazemzadeh,F.Bar및S.Narayanan,"2012년미국대통령선거주기의실시간트위터감정분석시스템",ACL 2012 시스템 데모 프로시딩, pp. 115-120, 2012.
16. L.Wang및J.Q.Gan,"트위터데이터분석을통한2017년프랑스선거예측",9회 컴퓨터 과학 및 전자 공학 회의(CEEC), IEEE, pp. 89-93, 2017.
17. M.Choy,M.L.Cheong,M.N.Laik및K.P.Shung,“2011년싱가포르대통령선거에대한인구조사보정 Twitter 데이터를 이용한 감성 분석,” arXiv preprint arXiv:1108.5520, 2011.
18. M.Skuza및A.Romanowski,"대규모분산환경에서의트위터데이터감성분석을이용한주식예측,"2015연계컴퓨터과학및 정보 시스템 회의(FedCSIS) 논문집, IEEE, pp. 1349-1354, 2015.
19. Y. Mao,W. Wei,B. Wang및 B.Liu, S&P500 주식과트위터 데이터를상관관계 분석,ACM 국제교차 학제적소셜 네트워크연구 핫토픽 워크숍논문집, pp. 69-72, 2012.
20. V.S.Pagolu,K.N.Reddy,G.Panda및B.Majhi,“트위터데이터를이용한주식시장동향예측을위한감성 분석,” 2016 국제 신호 처리, 통신, 전력 및 임베디드 시스템 회의(SCOPES) 논문집, IEEE, pp. 1345-1350, 2016.
21. X. Wang,M. S.Gerber 및D. E.Brown, “트위터게시물에서 추출한이벤트를 이용한자동 범죄예측,” 국제사회 계산,행동-문화 모델링및 예측회의, Springer,베를린, 하이델베르크, pp. 231-238, 2012.
22. Z.Miller,B.Dickinson,W.Deitrick,W.HuandA.H.Wang,Twitterspammerdetectionusingdata stream clustering, Information Sciences, 260, 64-73, 2014.
23. S.Munir,S.Wasi및S.I.Jami,"Urdu텍스트를위한토픽모델링접근법비교", 인디언 저널 오브 사이언스 앤 테크놀로지, 12, 45, 2019.
24. R.Vīksna,M.Kirikova및D.Kiopa,라트비아법적문서에서주제분석의활용탐구, 2018.
25. J.O'Neill,C.Robin,L.O'Brien및P.Buitelaar,입법문서에대한주제 모델링 분석, CEUR Workshop Proceedings, 2016.
26. C.P.George,D.Z.Wang,J.N.Wilson,L.M.Epstein,P.Garland및A.Suh,"설문조사에대한머신러닝 기반 주제 탐구 및 분류," 2012년 제11회 머신러닝 및 응용 국제 회의, Vol. 2, IEEE, pp. 7-12, 2012.
27. D.J.Hu,"텍스트,이미지및음악을위한잠재디리클레할당",캘리포니아대학교샌디에이고, 2009, 2013년 4월 26일 검색.
28. P.Xie,Y.Deng및E.Xing,"문서모델링을위한제한볼츠만머신다양화,"ACMSIGKDD국제지식 발견 및 데이터 마이닝 회의 회의록, pp. 1315-1324, 2015.
29. I.사토와H.나카가와,"Pitman-Yor과정을이용한파워로우토픽모델",ACMSIGKDD 국제 지식 발견 및 데이터 마이닝 회의 회의록, pp. 673-682, 2010.
30. A.Ahmed,L.Hong,A.Smola,"중첩된중국레스토랑프랜차이즈프로세스:사용자추적및문서모델링에 대한 응용", 국제 머신러닝 컨퍼런스, pp. 1426-1434, 2013.
31. L.AlSumait,D.Barbará및C.Domeniconi,“온라인LDA:텍스트스트림을위한적응형토픽모델및주제감지및추적에대한응용,” 2008 IEEE 국제 데이터 마이닝 회의, IEEE, pp. 3-12, 2008.
32. S.A.Curiskis,B.Drake,T.R.Osborn및P.J.Kennedy,두온라인소셜네트워크: 트위터와 레딧에서의 문서 클러스터링 및 주제 모델링 평가, 정보 처리 및 관리, 57(2), 102034, 2020.
33. M.CamprandK.Ježek,Comparingsemanticmodelsforevaluatingautomaticdocumentsummarization,In International Conference on Text, Speech, and Dialogue, Springer, Cham. pp. 252-260, 2015.
34. R. Das,M. Zaheerand C.Dyer, Gaussianlda fortopic modelswith wordembeddings, InProceedings ofthe 53rdAnnual Meetingof theAssociation forComputational Linguisticsand the7th International Joint Conference on Natural Language Processing, Volume 1: Long Papers, pp. 795-804, 2015.
35. R.Lin,S.Liu,M.Yang,M.Li,M.Zhou및S.Li,문서모델링을위한계층적순환신경망, 2015년 자연어 처리 실험 방법에 관한 회의 논문집, pp. 899-907, 2015.
36. Y.Liu,A.Niculescu-Mizil및W.Gryc,"주제-링크LDA:주제와저자커뮤니티의공동모델",제26회국제 머신러닝 회의 논문집, pp. 665-672, 2009.
37. Y. Gao,Y. Xu및 Y.Li, 정보필터링에서 문서모델링을 위한패턴 기반주제. IEEEKnowledge andData EngineeringTransactions, 27(6),1629-1642, 2014, http://dx.doi.org/10.1109/TKDE.2014.2384497.
38. 카글: NIPS 2015논문. 다음 링크에서확인할 수 있습니다: https://www.kaggle.com/benhamner/nips-2015-papers.
39. F.Rosner,A.Hinneburg,M.Röder,M.Nettling,andA.Both,"주제일관성측정방법평가,"arXivpreprint,arXiv:1403.6397,2014.
40. Q.Pleple,주제일관성으로주제모델평가하기,2013,다음링크에서확인가능:http://qpleple.com/topic-coherence-to-evaluate-topic-models,확인일: 2021년 1월 27일.
41. M.Röder,A.Both및A.Hinneburg,"주제일관성측정의공간탐구"2015년ACM국제 웹 검색 및 데이터 마이닝 회의 제8회 회의록, pp. 399-408.
42. A.드발과E.바나드,주제[모델의](http://researchspace.csir.co.za/dspace/bitstream/handle/10204/3016/de%20Waal1_2008.pdf?sequence=1&isAllowed=y)안정성 평가, http://researchspace.csir.co.za/dspace/bitstream/handle/10204/3016/de%20Waal1\_2008.pdf?sequence=1&isAllowed=y, 2008.
43. H.M.Wallach,I.Murray,R.Salakhutdinov및D.Mimno,“주제모델의평가방법,”제26회국제머신러닝회의논문집,pp. 1105-1112, 2009.
44. Q.Pleple,"주제모델평가방법:퍼플렉서티,"http://qpleple.com/perplexity-to-evaluate-topic-models/(2013),2021년1월28일 접속.
45. I. Anger및 C.Kittl, "트위터에서의영향력 측정",회의: 제11회지식 관리및 지식기술 국제회의 논문집,다음에서 이용가능: https://www.researchgate.net/publication/220866465\_Measuring\_influence\_on\_Twitter,2011, 2021년2월 15일접속, DOI: 10.1145/2024288.2024326.
46. F.Riquelme및P.Gonzalez-Cantergiani,"트위터에서사용자영향력측정:설문조사",정보처리및관리저널,52(5),Availableat:https://www.researchgate.net/publication/281427547\_Measuring\_user\_influence\_on\_Twitter\_A\_survey, 2016, 2021년 2월 15일 접속, DOI:10.1016/j.ipm.2016.04.003.

**콜린 N.우다노르의 사진과**생애에 대한정보는 출판시점에 이용할 수 없습니다.

**Onyinye E. Nweke의**사진과 생애에 대한정보는 출판 시점에는확인할 수 없습니다.

**조지 E. 오케르케, 사진및 생애에 대한 정보는**출판 당시 확인할 수 없습니다.

1. *(Corresponding author: Onyinye .E. Nweke).*

   The authors are with the Department of Computer Science, University of Nigeria, Nsukka Enugu State, Nigeria (e-mail: collins.udanor@unn.edu.ng; onyinye.nweke@unn.edu.ng; george.okereke@unn.edu.ng). [↑](#footnote-ref-1)