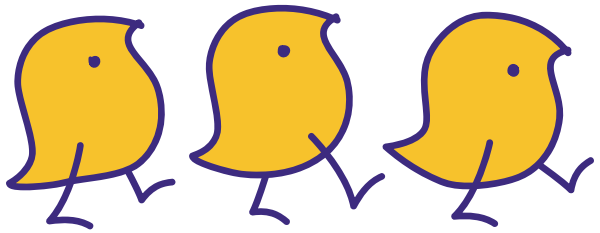
LangCon 2019

가짜뉴스 판별기 제작기

아주대학교 정건용 / 신민철

목차

1. Fake news(가짜뉴스)
2. 데이터 수집
3. 데이터 전처리
4. 데이터 분석
4. 결론
5. 한계점



Fake news(가짜뉴스)

LangCon 2019

가짜 뉴스의 정의

가짜 뉴스의 정의와 범위에 대해선 여러 의견이 있습니다.

언론사의 오보에서부터 인터넷 루머까지
넓은 스펙트럼 안에서 혼용되고 있습니다.

가짜 뉴스의 정의

가짜 뉴스?

- 정치, 경제적 이익을 위해 의도적으로 언론보도의 형식을 하고 유포된 거짓 정보

- 한국언론진흥재단

가짜 정보의 유형

- **허위정보(disinformation)**

  의도적으로 만들어진 허위정보 또는 오해를 부르는 정보

  대상을 속이기 위해 숙고해서 계산된 방식으로 퍼뜨리는 정보

- **오인정보(misinformation)**

  사실이 아님을 인지하지 못한 채 의도적 또는 비의도적으로 전파되는 정보

- **패러디(parodies)/풍자적**

  대상이 허구임을 인지할 수 있는 상태로 허위적 정보를 구성

- **거짓정보(hoax)**

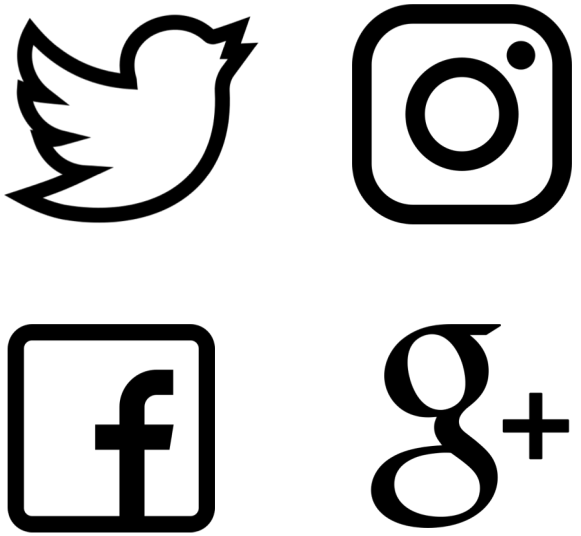
  진실을 가장해서 고의로 조작한 정보

  관찰이나 판단오류, 선의의 거짓말, 전설, 만우절 농담과 구별

- **루머/유언비어(rumor)**

  근거 없이 퍼지는 소문 등으로 정보의 불확실성이 주요 원인 중 하나

Mouth to Mouth



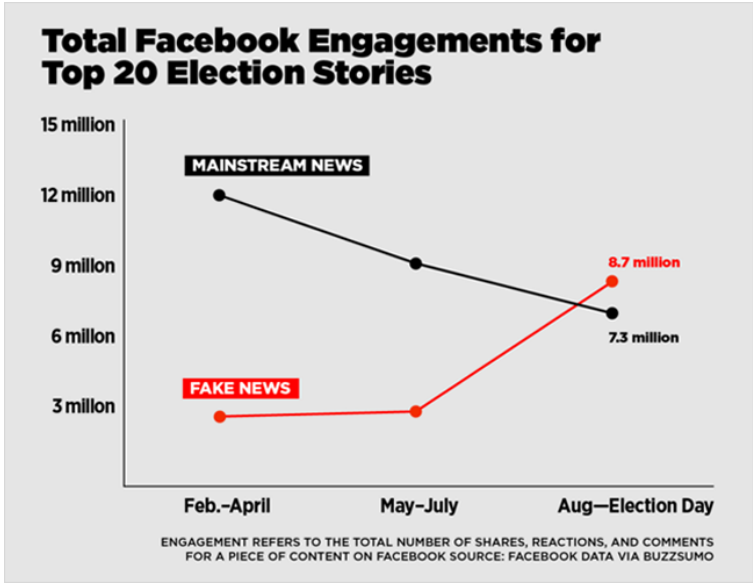
Media platform

가짜뉴스

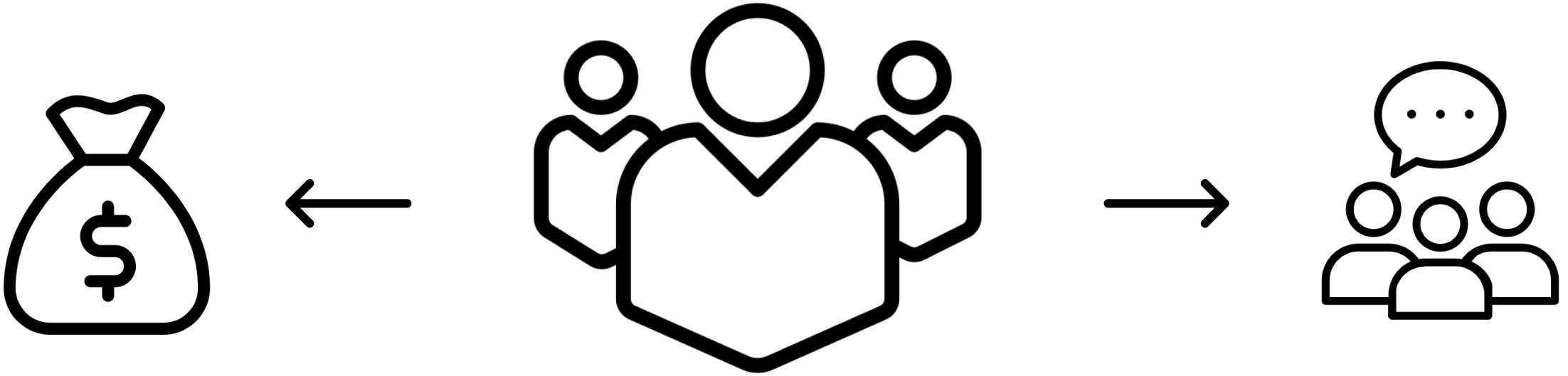
- **홍석천**

  경리단길 임대료 문제를 주된 이유로 가게를 그만둔다는 '이데일리'와 진행한 인터뷰가 왜곡되어 중앙일보에 보도

- 2016년 미국 대선기간중 8월에 가짜뉴스가 페이스북에서 주요 매체의 실제 뉴스보다 더 많은 반응을 이끌어 냄



- 주류언론뉴스와 가짜뉴스의 페이스북 반응 추이 〈출처 : 버즈피드〉

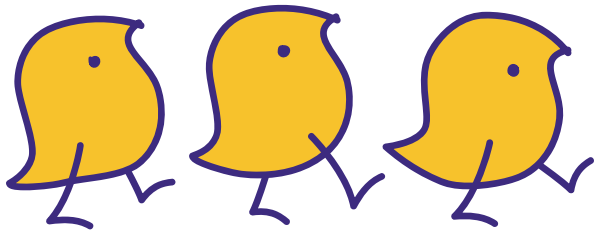


어떤 개인이나 집단의 목적달성

판별기의 목적

'사람들이 왜곡된 정보로 피해를 입거나

합리적인 의사결정을 방해하는 일이 일이 일어나지 않도록'



데이터 수집

LangCon 2019

데이터 수집

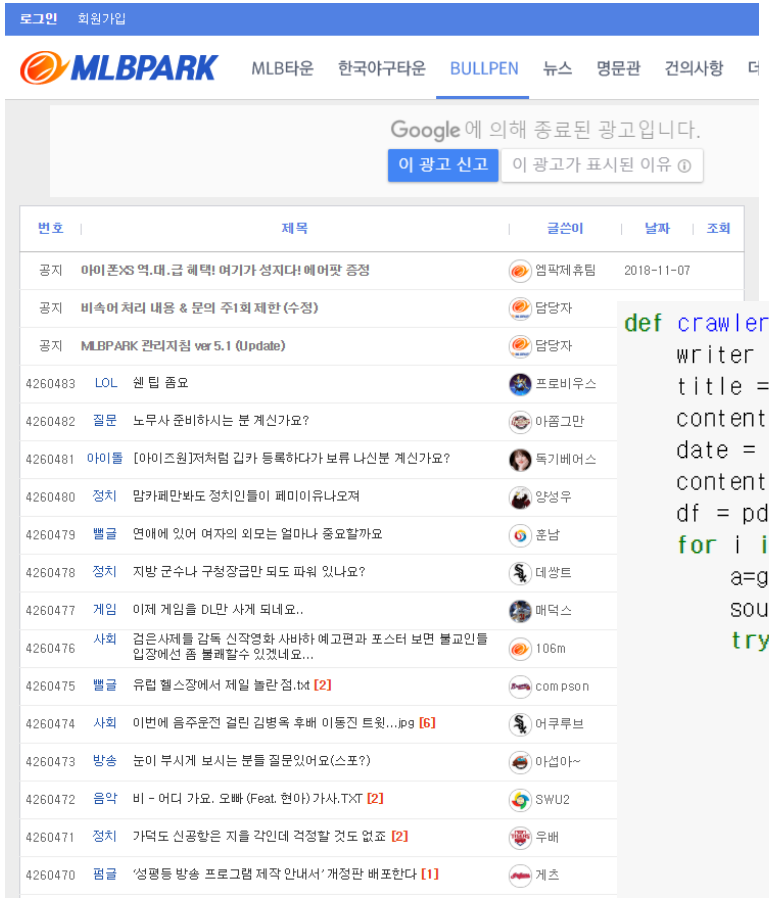
국내 커뮤니티 순위 집계 18.12.16-19.01.16

순위	사이트명	월방문자수(단위:만)
1	디씨	61.32
2	인벤	32.84
3	일베	21.82
4	펨코	19.51
5	루리	19.11
6	뽐뿌	15.21
7	클리앙	14.57
8	네판	13.88
9	엠팍	13.61
10	더쿠	12.37

사람들이 많은 커뮤니티 위주로 2가지 사이트 선정

데이터 수집

Language Conference 2019



자유게시판 성격의 MLBPARK BULLPEN 게시판 크롤링

```
def crawler_bon(subjects):
    writer = [] #글쓴이
    title = [] # 제목
    content = [] # 본문
    date = [] # 게시일
    content_type = [] #구분
    df = pd.DataFrame(columns=['구분' , '작성자', '작성일', '제목', '내용'])
    for i in subjects:
        a=get_html(i)
        soup = BeautifulSoup(a, 'html.parser')
        try:
            writer= soup.find_all("div", {"class":"text1"})
            content = soup.find_all("div", {"class":"ar_txt"})
            content_type = soup.find("div",{"class": "titles"}).find_all("span",{"class":"word"})
            title = soup.find_all("div",{"class": "titles"})
            title = title[0].text

            if (content_type[0].text=="아이돌" or content_type[0].text=="17금" or content_type[0].text=="스포츠" or content_type[0].text=="19
                title = title[3:len(title)]
            else :
                title = title[2:len(title)]

            date = soup.find("div",{"class": "text3"}).find_all("span", {"class":"val"})
            df = df.append(pd.Series([content_type[0].text, writer[0].text, date[0].text,title, content[0].text], index=df.columns ), ignore_
        except:
            pass
    return df
```

MLB PARK BULLPEN 게시판 게시글 내 댓글 크롤링

```
def crawler_daet(subjects):

    reply_content=[] #댓글본문
    reply_ID = [] # 댓글ID
    reply_date = [] # 댓글 시간
    reply_IP = [] # 댓글 IP
    df=pd.DataFrame()
    df = pd.DataFrame(columns=['본문제목','댓글ID' , '댓글시간', '댓글본문', '댓글IP'])

    for i in subjects:
        a=get_html(i)
        soup = BeautifulSoup(a, 'html.parser')
        content_type = soup.find("div",{"class": "titles"}).find_all("span", {"class":"word"})

        #글쓴이
        title = soup.find_all("div",{"class": "titles"})
        title = title[0].text
        if (content_type[0].text=="아이돌" or content_type[0].text=="17금" or content_type[0].text=="스포츠" or content_type[0].text=="19금"
            title = title[3:len(title)]
        else :
            title = title[2:len(title)]

        reply_ID = soup.find_all("span",{"class": "name"})
        reply_content = soup.find_all("span",{"class": "re_txt"})
        reply_date = soup.find("div",{"class": "reply_list"}).find_all("span", {"class":"date"})

        reply_IP = soup.find_all("span",{"class": "ip"})
        loop_num = len(reply_content)
        for j in range(loop_num):
            df = df.append(pd.Series([title, reply_ID[j].text, reply_date[j].text, reply_content[j].text,reply_IP[j].text], index=df.columns
    return df
```

데이터 수집

데이터 예시

	구분	작성자	작성일	제목	내용
0	밸글	에태몬	2019-02-12 09:55	박지성은 딱 레알선수의 이선수랑 일치해요.	루카스바스케스 이 선수랑 역할부터해서 비슷한 거 같은데 이분이 손흥민선수보다 잘하냐라...
1	질문	sicjwlss	2019-02-12 09:54	판교역 주변 주말에도 차 많이 막히나요?	주말에 여자친구 내려줘야 하는데.... 차막히는 곳은 힘들어서 ㅠ 판교역 부근 차로...
2	사회	하이텐	2019-02-12 09:54	유교 탈레반 국가 대한민국...jpg	싹ㅡ다 금ㅡ지
3	사회	감베리니	2019-02-12 09:54	여자 173에 65kg 어느정도 체형인가요?	이영자 체형인가요? 아님 보기좋은 정도인가요? 중고등학교때 육상도 했다고 하던데..
4	스포츠	redd	2019-02-12 09:54	배구 플옵은 몇 강인가요?	3위까지 플옵 가는건가요 4위까지 플옵 가는건가요?
5	짤방	SECG	2019-02-12 09:53	공항에서 빅뱅인척 하는 일반인.gif	태양승리대성지디탑ㅎㅎㅎㅎㅎㅎ
6	정치	혼다미오	2019-02-12 09:53	https 차단 반대 청원 링크.jpg	https://www1.president.go.kr/petitions/522031?...

2월 중의 커뮤니티 사이트 본문/댓글 관련 데이터 수집

데이터 수집

**가짜뉴스는 특정 커뮤니티나 유사 언론 사이트를 숙주 삼아 생산되고,
페이스북과 카카오톡 등 소셜미디어를 통해 확산되는 전형적 메커니즘을 따름**

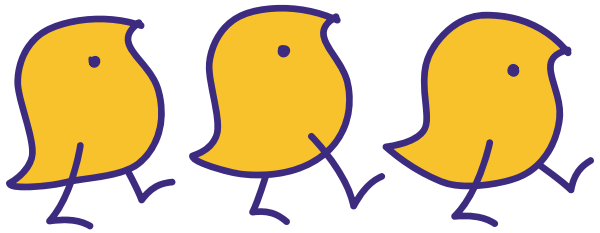
인터넷 서비스 중 가짜뉴스를 받아본 경로 (%)	
카카오톡, 라인 등 모바일 메신저	39.7
페이스북, 트위터 등 소셜플랫폼	27.7
인터넷 카페/커뮤니티, 블로그	24.3
유튜브, 아프리카TV 등	4.6
가짜 뉴스 사이트	3.7

한국언론진흥재단

Eco-Chamber

닫힌 시스템 안에서 어떤 정보나 아이디어, 신념들이 교환 될수록 증폭되고 강화되며, 그와 다르거나 반하는 시각들은 검열되거나 불허 혹은 축소 반영 되는 상황

가짜 뉴스의 신뢰성 ↑

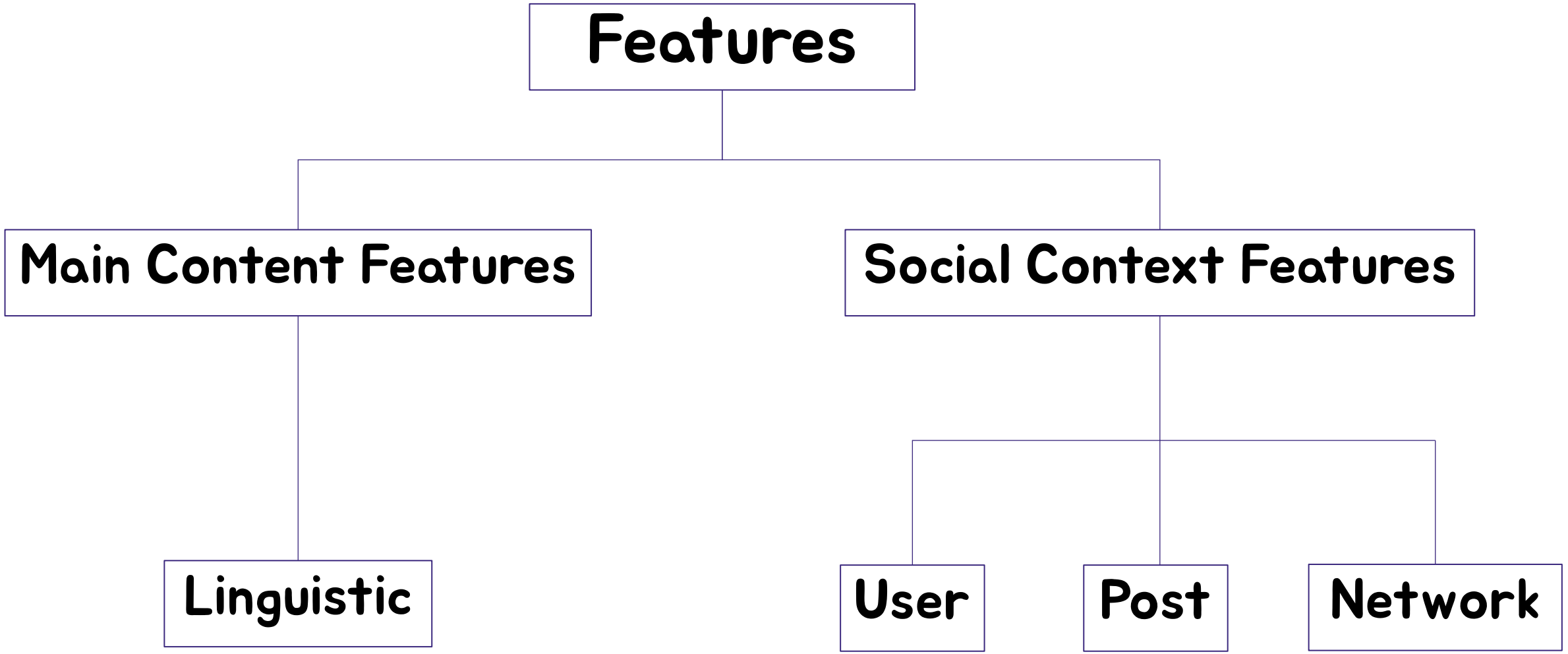


- Feature Extraction

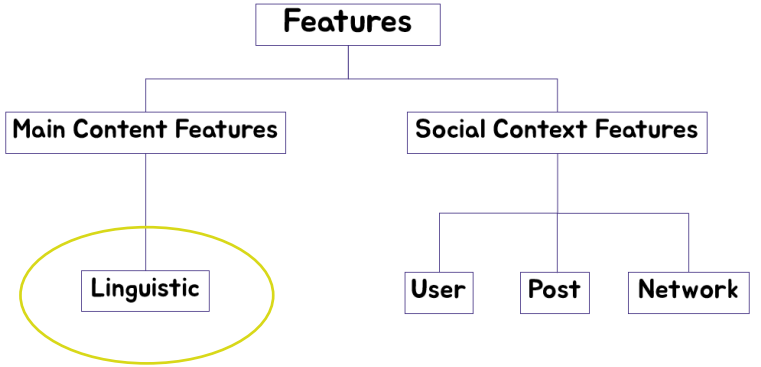
데이터 전처리

LangCon 2019

Feature Extraction



Linguistic-based Feature



Since fake news pieces are intentionally created for financial or political gain rather than to report objective claims, they often contain opinionated and inflammatory language, crafted as 'clickbait'

It is reasonable to exploit linguistic features that capture the different writing styles and sensational headlines to detect fake.

- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

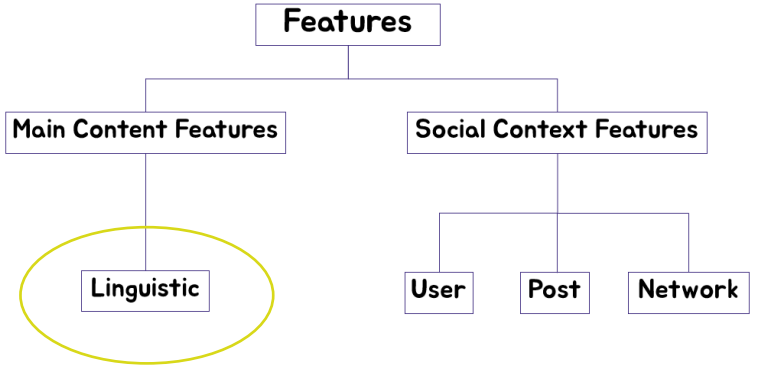
Feature Extraction

Fake News일수록 객관적인 주장보다 정치/경제관련 글과

독선적이고 선정적인 단어가 포함되는 경향이 있기 때문에

언어적인 특징을 이용해 Fake News를 찾아내는 것이 타당하다.

Linguistic-based Feature



1. 내용
   - 본문 내의 텍스트

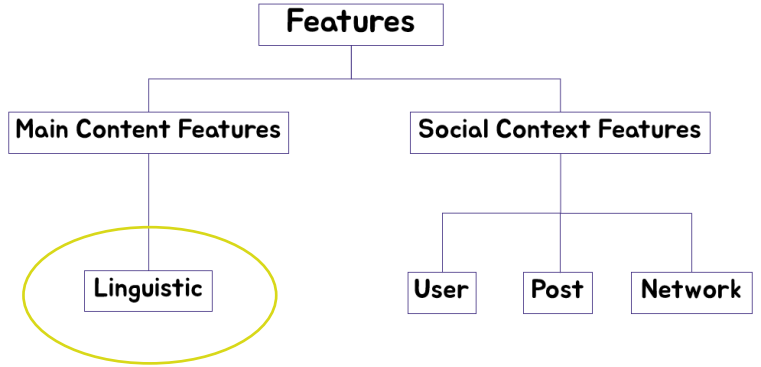
Feature Extraction

SoyNLP 이용하여 Nomarlizing

Khaiii(카카오 형태소 분석기) 사용하여 Tokenizing

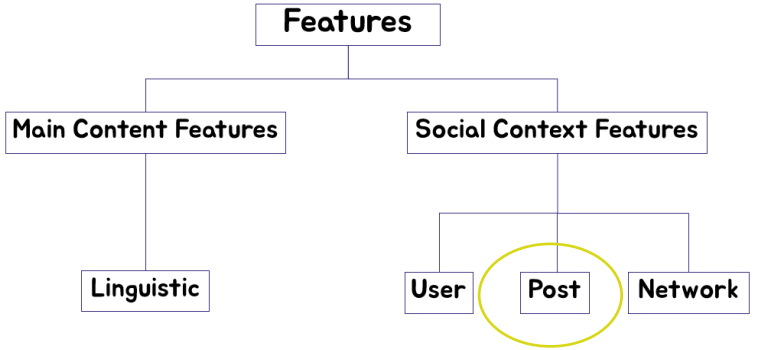
FastText 사용하여 Vector Embedding

각 게시글 텍스트에 대해 TF-IDF 벡터를 산출하여 FastText Vector를 적용



	0	1	2	3	4	5	6	7	
주말에	0.062367	0.019849	0.009015	0.014948	-0.039134	0.022025	0.150986	-0.137575	-0.1521
힘들어서	0.070205	0.019947	0.018845	0.023239	-0.039603	0.029865	0.142906	-0.144213	-0.16681
전차의	0.059685	0.019623	0.008162	0.019539	-0.039639	0.016294	0.148264	-0.138310	-0.1554
기술들	0.055344	0.021958	-0.003383	0.021755	-0.047207	0.009930	0.172400	-0.146275	-0.1707
파워팩들	0.067067	0.018215	0.006144	0.019930	-0.044244	0.024988	0.162384	-0.149663	-0.1720
취재	0.061728	0.015738	0.002706	0.019984	-0.040561	0.016560	0.138655	-0.129080	-0.1508
영하	0.037188	0.012687	0.002120	0.011531	-0.029683	0.020032	0.099161	-0.091950	-0.1039
자동으로	0.070288	0.018142	0.011265	0.019555	-0.049936	0.029762	0.166104	-0.164003	-0.1913

Post-based Feature



People express their emotions or opinions towards fake news through social media posts, such as skeptical opinions, sensational reactions, etc.

Thus, it is reasonable to extract post-based features to help find potential fake news via reactions from the general public as expressed in posts.

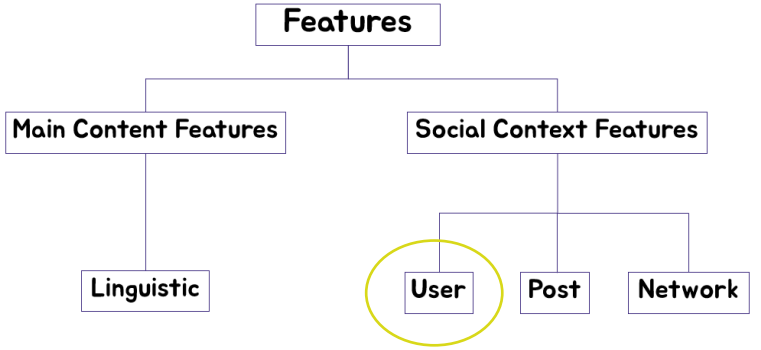
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

Feature Extraction

사람들은 회의적인 의견, 선정적인 반응 등 SNS를 통해 가짜 뉴스에 대한 감정이나 의견을 표현한다.

따라서 포스트에 표현된 일반 대중의 반응을 통해 잠재적인 가짜 뉴스를 찾는데 도움이 되는 포스트 기반 특징을 추출하는 것이 타당하다

User-based Feature



Fake news pieces are likely to be created and spread by non-human accounts, such as social bots or cyborgs.

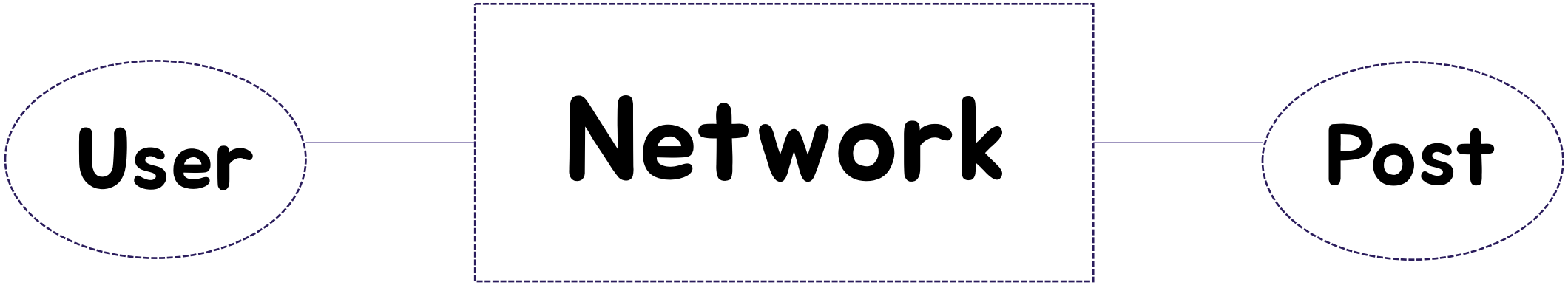
Thus, capturing users' profiles and characteristics by user-based features can provide useful information for fake news detection.

- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

Feature Extraction

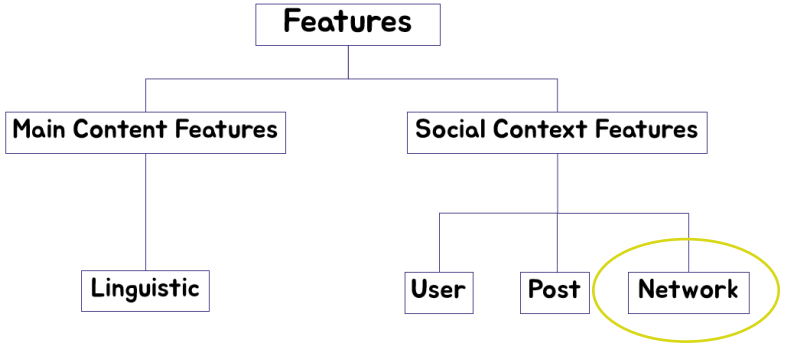
사람이 아닌 봇에 의해 가짜뉴스가 생성될 수 있기 때문에

User의 Profile인 User Based Feature를 사용하여 Fake News를 찾아내는 것이 타당하다.



User와 Post 각각의 Feature를 찾는것 보다
둘사이의 네트워크에 주목

Network-based Feature



Users form different networks on social media in terms of interests, topics, and relations.

Fake news dissemination processes tend to form an echo chamber cycle, highlighting the value of extracting network-based features to represent these types of network patterns for fake news detection.

- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

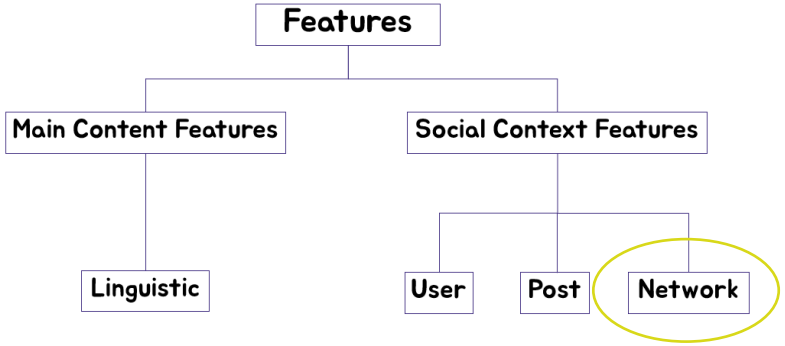
Feature Extraction

Language Conference 2019

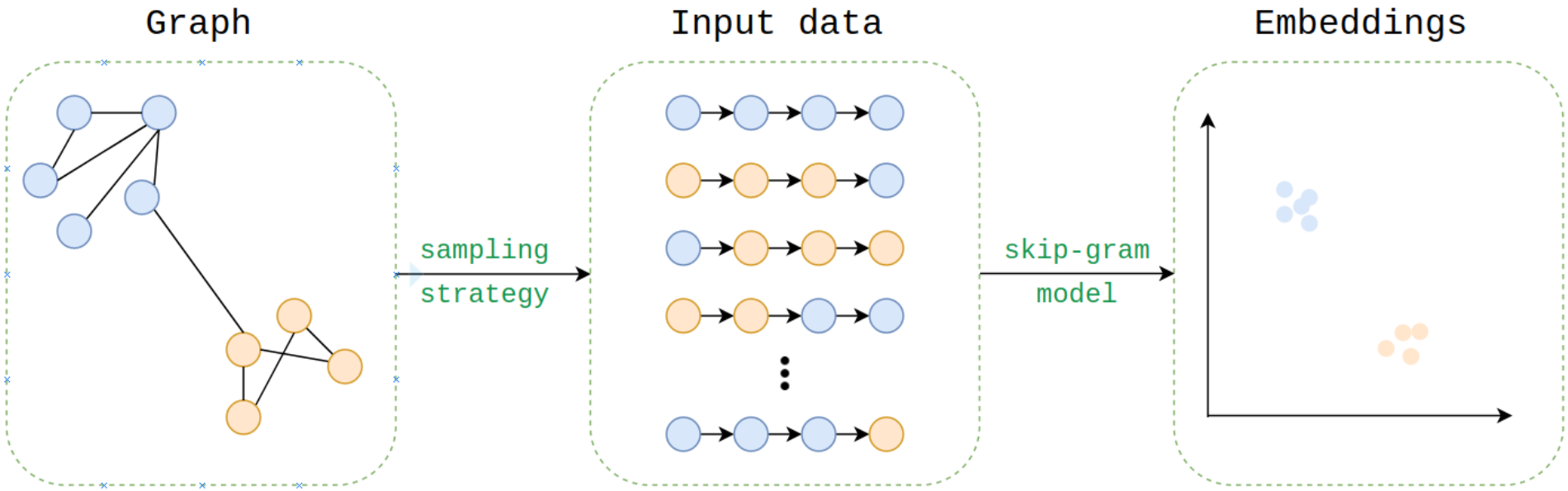
사용자는 관심사, 주제 및 관계 측면에서 소셜 미디어에서 서로 다른 네트워크를 형성한다.

가짜 뉴스 검출을 위해 이러한 유형의 네트워크 패턴을 추출하는 것이 타당하다.

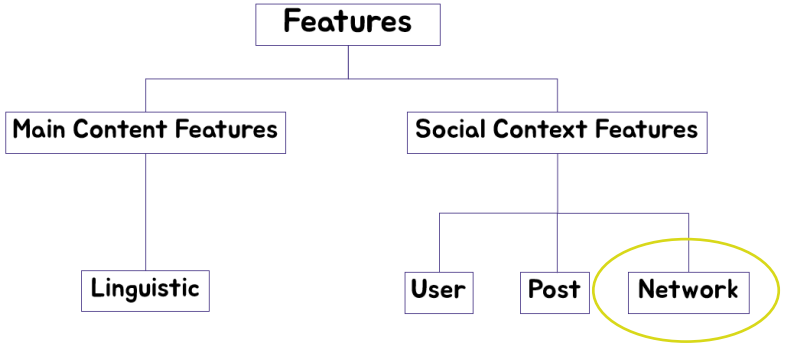
Network-based Feature



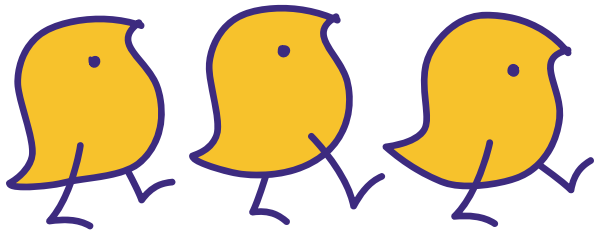
Node2Vec



Network-based Feature



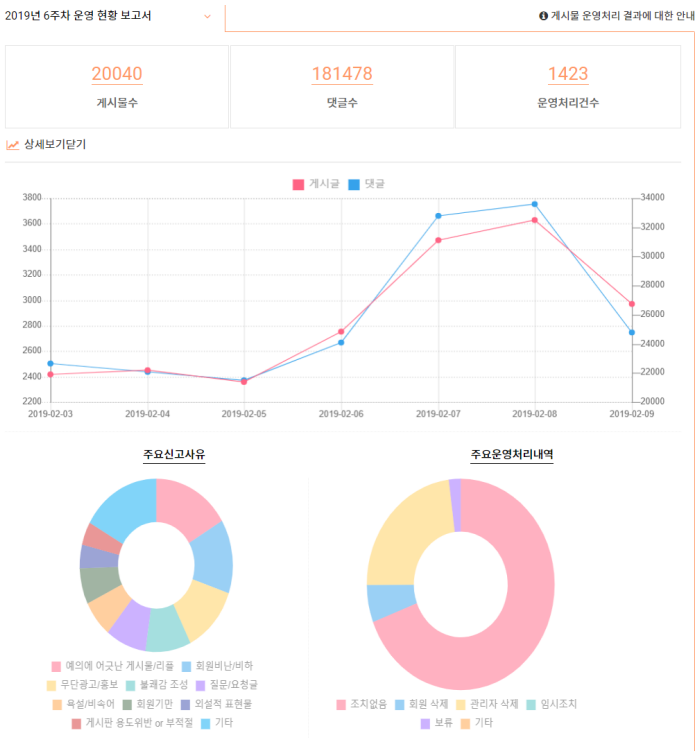
1. Network Feature
   - 작성자 ~ 댓글 작성자 간의 관계를 Interaction으로 가정한 후 Network 관계를 수집하여 Node2Vec으로 벡터화



데이터 분석

LangCon 2019

Data Labeling



- 커뮤니티 운영 현황 보고서를 참고하여

사람들이 신고한 게시글과 자진삭제한 게시글을

가짜뉴스라고 가정

모델

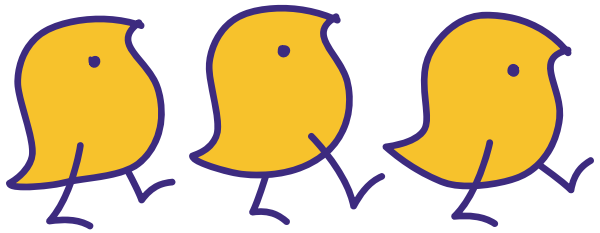
XGBoost

Grid Search

평가

6000여개의 게시물 데이터

Cross Validation



결론

예측

XGBoost Classification

In [36]:
```
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
```

In [37]:
```
X1 = mlbpark_dataset.iloc[0:2507,5:-2]
Y1 = mlbpark_dataset.iloc[0:2507,-1]
```

In [38]:
```
X2 = mlbpark_dataset.iloc[2507:5391,5:-2]
Y2 = mlbpark_dataset.iloc[2507:5391,-1]
```

In [39]:
```
cv_params = {'learning_rate': [0.1, 0.01], 'subsample': [0.7,0.8,0.9]}
ind_params = {'n_estimators': 1000, 'seed':0, 'colsample_bytree': 0.8,
              'objective': 'binary:logistic', 'max_depth': 3, 'min_child_weight': 1, 'tree

optimized_GBM = GridSearchCV(XGBClassifier(**ind_params),
                             cv_params,
                              scoring = 'accuracy', cv = 5, n_jobs = -1, verbose=10)
```

In [40]:
```
result = optimized_GBM.fit(X1, Y1)
```

```
Fitting 5 folds for each of 6 candidates, totalling 30 fits
[CV] learning_rate=0.1, subsample=0.7 ................................
[CV] learning_rate=0.1, subsample=0.7 ................................
```

- X1,Y1과 X2,Y2를 나눠서 Validation Set을 설정하였으며 Hyper parameter 최적화를 위해 Grid Search 하였습니다.

예측

```
In [41]: result.grid_scores_
```

```
/opt/conda/lib/python3.6/site-packages/sklearn/model_selection/_search.py:762: DeprecationWarning: The grid_scores_ attribute was deprecated in version 0.18 in favor of the more elaborate cv_results_ attribute. The grid_scores_ attribute will not be available from 0.20
  DeprecationWarning)
```

```
Out[41]: [mean: 0.91424, std: 0.01034, params: {'learning_rate': 0.1, 'subsample': 0.7},
 mean: 0.91424, std: 0.01087, params: {'learning_rate': 0.1, 'subsample': 0.8},
 mean: 0.91663, std: 0.00948, params: {'learning_rate': 0.1, 'subsample': 0.9},
 mean: 0.92461, std: 0.00606, params: {'learning_rate': 0.01, 'subsample': 0.7},
 mean: 0.92621, std: 0.00464, params: {'learning_rate': 0.01, 'subsample': 0.8},
 mean: 0.92541, std: 0.00554, params: {'learning_rate': 0.01, 'subsample': 0.9}]
```

Predict

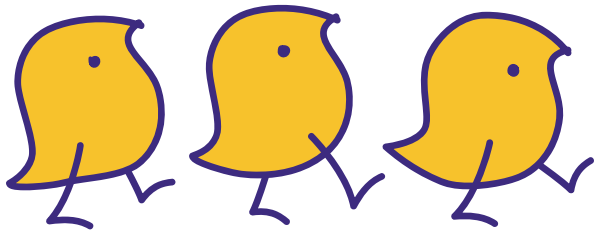
```
In [48]: test_probs = result.predict_proba(X2)[:,1]
```

```
In [50]: predictions = [round(value) for value in test_probs]
```

```
In [53]: from sklearn.metrics import accuracy_score
         accuracy = accuracy_score(Y2,predictions)
         accuracy
```

```
Out[53]: 0.9733009708737864
```

- 모델정확도는 92.5%
- Validation의 Accuracy는 97.3%



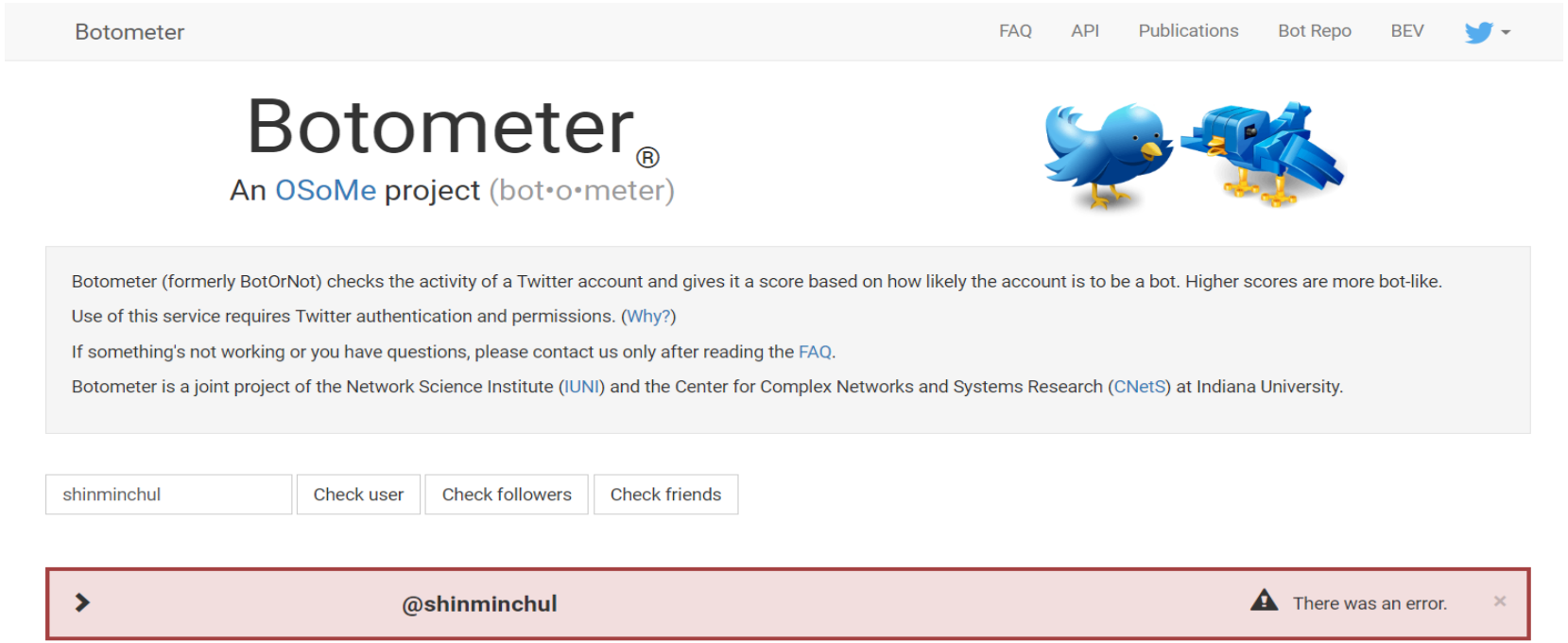
한계점

한계

- 적은 양의 데이터

- 추가적인 feature
  - 댓글 작성자들 사이의 NetWork Feature
  - 작성자의 글 중 작성자의 댓글 / 전체 댓글 비율
  - 댓글 작성자들의 Text를 Feature로 사용

Future Work



Fake News로 의심되는 게시글의 URL이나

유저의 정보를 집어 넣었을때 결과를 반환해주는 웹사이트