



# **CLASSIFYING FAKE NEWS ARTICLES USING NATURAL LANGUAGE PROCESSING TO IDENTIFY IN-ARTICLE ATTRIBUTION AS A SUPERVISED LEARNING ESTIMATOR**

**Sannidhiraju N V Satya Lakshmi Sujitha** (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

**Dr. I. R. Krishnam Raju**, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

## **ABSTRACT**

Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called ‘fake news’ is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Textblob, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in a Bayesian machine learning system as a key feature to estimate the likelihood that a news article is fake. The resultant process precision is 63.333% effective at assessing the likelihood that an article with quotes is fake. This process is called influence mining and this novel technique is presented as a method that can be used to enable fake news and even propaganda detection. In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system.

## **1. INTRODUCTION**

Intentionally deceptive content presented under the guise of legitimate journalism (or ‘fake news,’ as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

## **2. LITERATURE SURVEY**

### **1) When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism**

**AUTHORS:** M. Balmas

This research assesses possible associations between viewing fake news (i.e., political satire) and attitudes of inefficacy, alienation, and cynicism toward political candidates. Using survey



data collected during the 2006 Israeli election campaign, the study provides evidence for an indirect positive effect of fake news viewing in fostering the feelings of inefficacy, alienation, and cynicism, through the mediator variable of perceived realism of fake news. Within this process, hard news viewing serves as a moderator of the association between viewing fake news and their perceived realism. It was also demonstrated that perceived realism of fake news is stronger among individuals with high exposure to fake news and low exposure to hard news than among those with high exposure to both fake and hard news. Overall, this study contributes to the scientific knowledge regarding the influence of the interaction between various types of media use on political effects.

## **2) Miley, CNN and The Onion**

**AUTHORS: D. Berkowitz and D. A. Schwartz**

Following a twerk-heavy performance by Miley Cyrus on the Video Music Awards program, CNN featured the story on the top of its website. The Onion—a fake-news organization—then ran a satirical column purporting to be by CNN's Web editor explaining this decision. Through textual analysis, this paper demonstrates how a Fifth Estate comprised of bloggers, columnists and fake-news organizations worked to relocate mainstream journalism back to within its professional boundaries.

## **3) The Impact of Real News about “Fake News” ’ : Intertextual Processes and Political Satire**

**AUTHORS: P. R. Brewer, D. G. Young, and M. Morreale**

This study builds on research about political humor, press metacoverage, and intertextuality to examine the effects of news coverage about political satire on audience members. The analysis uses experimental data to test whether news coverage of Stephen Colbert's Super PAC influenced knowledge and opinion regarding *Citizens United*, as well as political trust and internal political efficacy. It also tests whether such effects depended on previous exposure to *The Colbert Report* (Colbert's satirical television show) and traditional news. Results indicate that exposure to news coverage of satire can influence knowledge, opinion, and political trust. Additionally, regular satire viewers may experience stronger effects on opinion, as well as increased internal efficacy, when consuming news coverage about issues previously highlighted in satire programming.

## **4) Stopping Fake News**

**AUTHORS: M. Haigh, T. Haigh, and N. I. Kozak**

Social media is acting as a double-edged sword for universe in a way of consuming news. On one side, its ease of access, popularity and low cost distribution channel lead people to gain news from social media. On other side, it is also acting as a source of spread of 'fake news'. The extensive spread of fake news on social media, websites are impacting society negatively. This makes extremely important to combat the spread of fake news and to aware the society. In this paper, we offer a review which lists out the sources of fake news, its types, generation, motivation and examples. Also, some approaches are suggested to spot and stop fake news spread.



## 5) With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity"

**AUTHORS:** R. Marchi

This article examines the news behaviors and attitudes of teenagers, an understudied demographic in the research on youth and news media. Based on interviews with 61 racially diverse high school students, it discusses how adolescents become informed about current events and why they prefer certain news formats to others. The results reveal changing ways news information is being accessed, new attitudes about what it means to be informed, and a youth preference for opinionated rather than objective news. This does not indicate that young people disregard the basic ideals of professional journalism but, rather, that they desire more authentic renderings of them.

### **3. IMPLEMENTATION**

#### **MODULES:**

- ❖ Social Media Mining System Construction
- ❖ User Topical Package Model Mining
- ❖ Route Package Mining
- ❖ Travel sequence recommendation

#### **MODULES DESCRIPTION:**

##### **Social Media Mining System Construction**

- ❖ In the first module we develop the system for the evaluation of our proposed model and thus make the system construction module with social media mining system.
- ❖ Our topic package space is the extension of textual descriptions of topics such as ODP. We use the topical package space to measure the similarity of the user topical model package (user package) and the route topical model package (route package). In our paper, we construct the topical package space by the combination of two social media: travelogues and community-contribute photos. To construct topical package space, travelogues are used to mine representative tags, distribution of cost and visiting time of each topic, while community-contributed photos are used to mine distribution of visiting time of each topic.
- ❖ The reasons for using the combination of social media are (1) travelogues are more comprehensive to describe a location than the tags with the photos which are with so many noises; (2) it is difficult to mine a user's consumption capability and the cost of POIs directly by the photos or the tags with the photos; (3) to season, although both media could offer correct visiting season information of POIs, the number of photos of a POI is far larger than the number of travelogues. (4) the time difference between where the user lives and the "data taken" of community contributed photos of where he or she visits make the taken time inaccurate.

##### **User Topical Package Model Mining**

- ❖ User topical package model (user package) is learnt from mapping the tags of user's photos to topical package space. It contains user topical interest distribution



(U), user consumption capability (U), preferred travel time distribution (U) and preferred travel season distribution .

- ❖ In this module, we introduce how to extract the user package, which contains user topical interest distribution, user consumption capability distribution, preferred travel time distribution and preferred travel season distribution.
- ❖ First we introduce user's topical interest mining from mapping user's tags to the topical package space. Then, we introduce how to get topical space mapping method.
- ❖ We map the textual description (tags) of user's community photos to the topical package space to present the user's travel preference of different topics, which is defined as user topical interest distribution. We assume that if a user's tags appear frequently in one topic and less in others, the user has a higher interest towards this topic.
- ❖ We use the cost distributions of the all the topics and distribution of use's topical interest to present a user's consumption capability. If a user usually takes part in luxurious activities like Golf and Spas, his consumption capability is very likely to be. If a user usually takes part in some cheap things, his consumption capability is likely to be low, and we tend not to recommend him luxurious topics.

### **Route Package Mining**

- ❖ Route topical package model (route package) is learnt from mapping the travelogues related to the POIs on the route to topical package space. It contains route topical interest, route's cost distribution, route's time distribution and season distribution.
- ❖ To save the online computing time, we mine travel routes and the attribute of the routes offline. After mining POIs, to construct travel routes, we analyze the spatio-temporal structure of the POIs among travelers' records.
- ❖ We construct the spatio-temporal structure of the POIs according to the "data taken". POI with the earlier timestamp is defined as the "in". POI with a later timestamp, on the contrary, is defined as "out". Then we count the times of "in" and "out" from POI to others by the records of all the users after filtering. A greedy algorithm is then applied to find the time sequence of these POIs. Thus, we finish famous routes mining and obtain famous routes of each city.

### **Travel sequence recommendation**

- ❖ After mining user package and route package, in this module, we develop our travel routes recommendation module. It contains two main steps: (1) routes ranking according to the similarity between user package and routes packages, and (2) route optimizing according to similar social users' records.
- ❖ After POI and route ranking module, we get a set of ranked routes. Here, we further describe the optimization of top ranked routes according to social similar users' travel records. Firstly, we introduce how to mine social similar users and their travel records. Then we introduce how to optimize the roads by social users' travel records.

3 techniques will be used to calculate score

- 1) Source: any person who is writing news will give his name or a person name on which he writing articles



- 2) CUE: using this we will extract VERBS or VERBS phrases, if news is real then it will have verb types of words
- 3) Quotes: all articles will be on some topics and person will describe that topic name under quotes. So we will look for quotes in articles to determine fake or real news.

## Document examples

"When "Mitt Romney" was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it."

In above sentence quotes are there and it's talking about 'Mitt Romney' and it's contains some verbs such as 'was, didn't, slow, cut'. By analysing above 3 features from articles we can come to the conclusion whether news is FAKE or REAL.

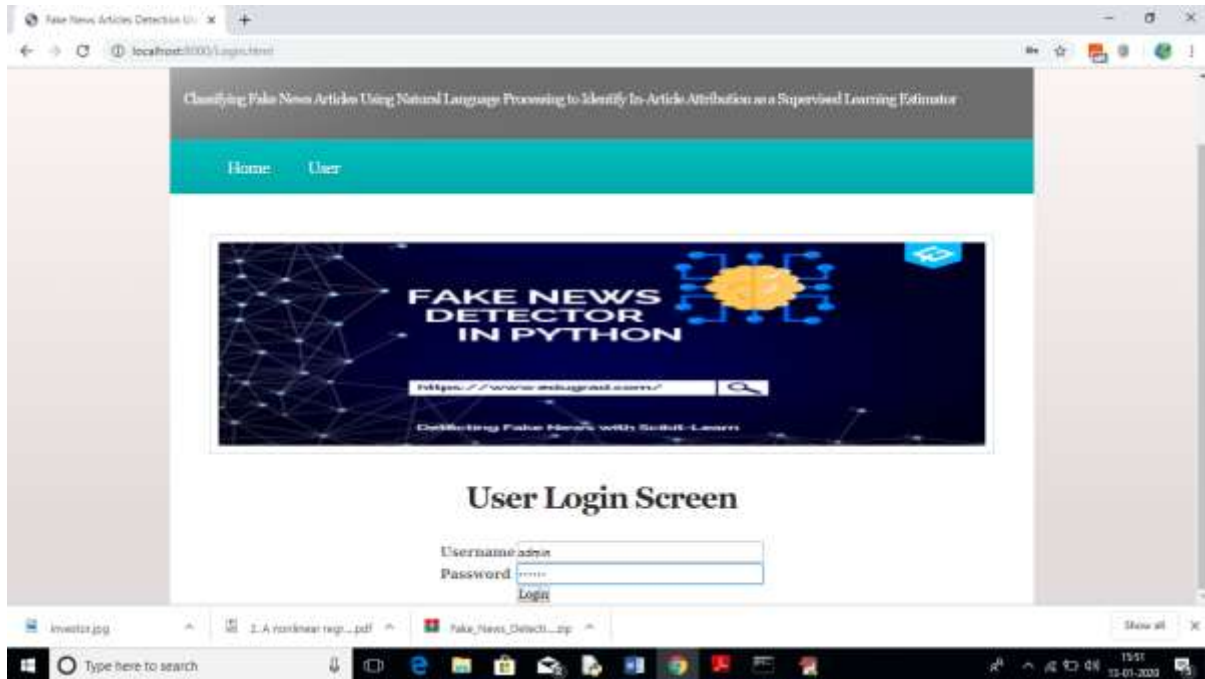
All FAKE peoples will not write such statements in their articles so we can detect by applying this techniques.

To implement this project we are using 'News' dataset and then by applying above technique we can detect whether this news are fake or real. This dataset I kept inside dataset folder. Upload this dataset when you are running application.

To run this project deploy 'FakeNews' folder on 'django' python web server and then start server and run in any web browser. After running code in web browser will get below page. Screen shots



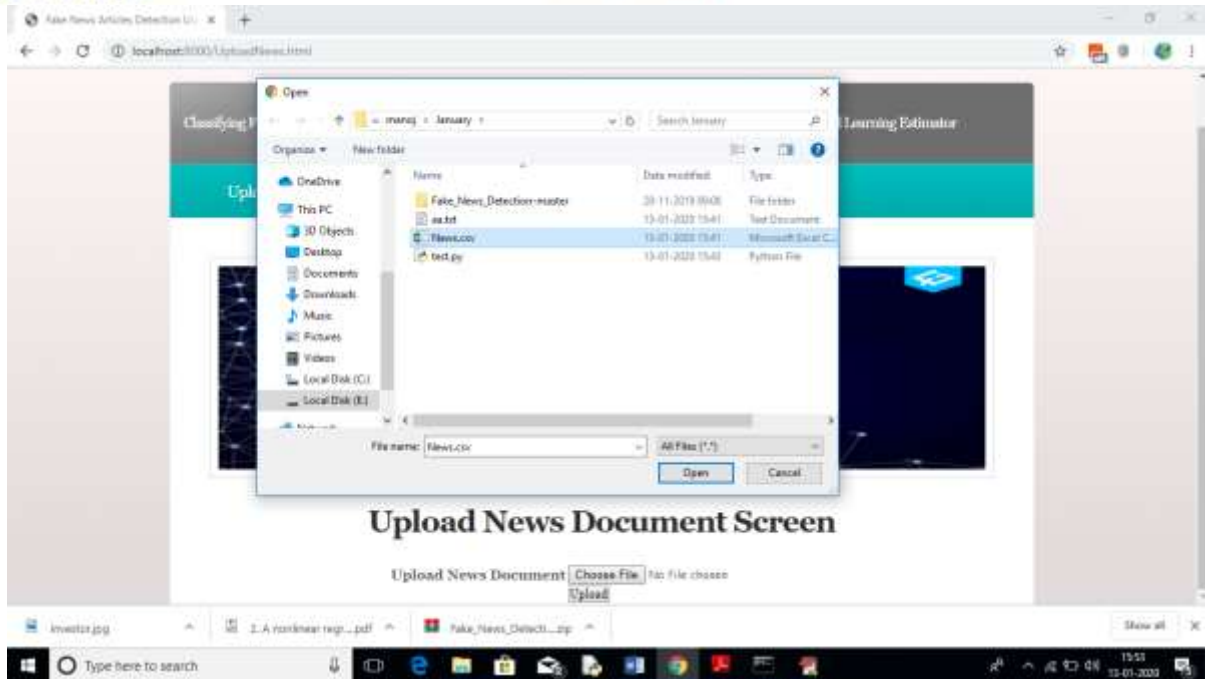
In above screen click on 'User' link to get below screen



In above screen enter username and password as 'admin' and then click on 'Login' button to get below screen



In above screen click on 'Upload News Articles' link to upload news document



Upload News Document Screen

In above screen I am uploading 'News.csv' file which contains 150 news paragraphs. After uploading news will get below screen



Upload News Document Screen

In above screen news file uploaded successfully, now click on 'Run Fake News Detector Algorithm' link to calculate Fake News Detection algorithm score and based on score and naïve bayes algorithm we will get result.



News Text	Detection Result	Fake Rank Score
Says the Anties List political group supports third-trimester abortions on demand.	Fake News	0.8333333333333333
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bush's administration.	Real News	2.142857142857143
"Hillary Clinton agrees with John McCain" by voting to give George Bush the benefit of the doubt on Iran.	Real News	3.076923076923077
Health care reform legislation is likely to mandate fine sex change surgeries.	Fake News	0.792307692307693
The economic turnaround started at the end of my term.	Real News	0.909090909090909
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (T/W) faculty fired during the last two decades.	Real News	1.333333333333333
Jim Dornan has not lived in the district he represents for years now.	Real News	2.142857142857143
I'm the only person on this stage who has worked actively just last year passing, along with Ross Feingold, some of the toughest ethics reform since Watergate."	Real News	1.515151515151515
"However, it took \$49.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific."	Real News	2.142857142857143
Says GOP primary opponents Glenn Grothman and Joe Leibham cast a compromise vote that cost \$788 million in higher electricity costs.	Real News	2.1739130434782608
For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin."	Fake News	0.8
"Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty."	Real News	1.5
"When Mitt Romney was governor of Massachusetts, we didn't just slow the rate of growth of our government, we actually cut it."	Real News	2.222222222222222
The economy bled \$24 billion due to the government shutdown.	Fake News	0.833333333333333
Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended.	Real News	2.1052631578947367
"In this last election in November, ... 63 percent of the American people chose not to vote. ... So percent of young people, (and) 75 percent of low-income workers chose not to vote."	Real News	0.973609736097361

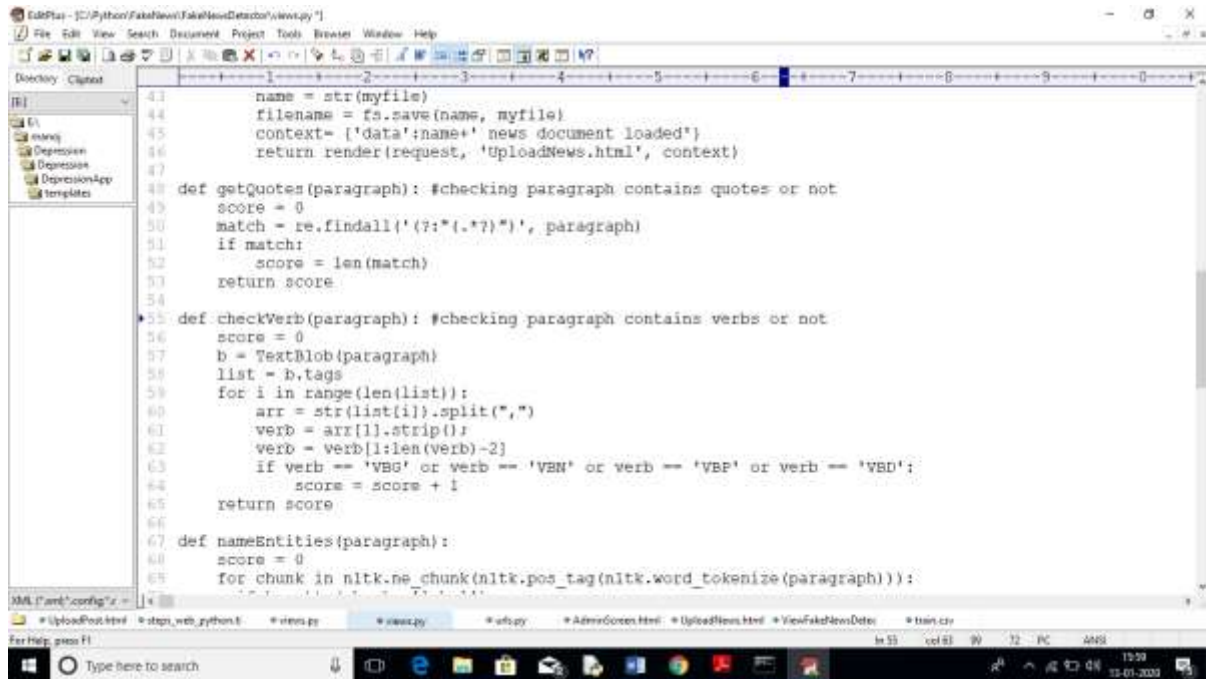
In above screen first column contains news text and second column is the result value as 'fake or real' and third column contains score. If score greater > 0.90 then I am considering news as REAL otherwise fake.

Some neighborhood schools are closing.	Real News	0.333333333333333
He told gay organizers in Massachusetts he would be a stronger advocate for special rights than even Ted Kennedy.	Real News	1.5
"The years that I was speaker, the Florida House consistently offered leaner budgets than the governor offered."	Real News	2.380952380952381
"We are already almost halfway to our 2010 goal of creating 700,000 new jobs in seven years."	Real News	1.5
Says the U.S. Supreme Court found that Social Security is not guaranteed.	Real News	1.8461538461538463
Says Michael Bennett wants to close Guantanamo Bay prison and bring terrorists right here to Colorado.	Real News	2.0000000000000005
Oregonians have an amazing no-cost way to fight abortion with free political donations	Fake News	2.7692307692307693
The president said he's going to bring in 250,000 (Syrian and Iraqi) refugees into this country.	Real News	2.380952380952381
Research shows that a vast majority of arriving immigrants today come here because they believe that government is the source of prosperity, and that's what they support."	Real News	1.6129032258064515
Newt Gingrich's immigration plan offers a new doorway to amnesty.	Real News	1.6181818181818183
Mr. Caprio is a career politician who has never worked in the private sector.	Real News	0.0
"In Rhode Island, 9 percent of workers use the states temporary disability insurance program each year while in New Jersey, the rate is only 3 percent."	Real News	1.3903225806451613
"In just 17 years, spending for Social Security, federal health care and interest on the debt will exceed ALL tax revenue!"	Fake News	0.7992307692307693
President Obama took more money from Wall Street in the 2008 campaign than anybody ever had.	Real News	2.3529411764705883
Donald Trump has said nuclear proliferation is OK.	Real News	0.333333333333333
Hillary Clinton has taken over \$800,000 from lobbyists."	Real News	2.5
Barack Obama has never even worked in business.	Real News	0.333333333333333
Says the Arizona immigration law expressly bans racial profiling.	Real News	1.0
Says Gov. Rick Perry has been begging for the federal government to send the Coast Guard to patrol two lakes on the U.S.-Mexico border.	Real News	1.0230769230769231
"On the VA: Over 300,000 veterans have died waiting for care."	Real News	2.0000000000000005

For all 150 news text articles we got result as fake or real.

See below screen shots of code calculating quotes, name entity and verbs from news paragraphs





```
41 name = str(myfile)
42 filename = fs.save(name, myfile)
43 context= ['data':name+' news document loaded']
44 return render(request, 'UploadNews.html', context)
45
46 def getQuotes(paragraph): #checking paragraph contains quotes or not
47 score = 0
48 match = re.findall('(?:".*?")', paragraph)
49 if match:
50 score = len(match)
51 return score
52
53 def checkVerb(paragraph): #checking paragraph contains verbs or not
54 score = 0
55 b = TextBlob(paragraph)
56 list = b.tags
57 for i in range(len(list)):
58 arr = str(list[i]).split(",")
59 verb = arr[1].strip()
60 verb = verb[1:len(verb)-2]
61 if verb == 'VBS' or verb == 'VBN' or verb == 'VBP' or verb == 'VBD':
62 score = score + 1
63 return score
64
65 def nameEntities(paragraph):
66 score = 0
67 for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(paragraph))):
```

## 4. CONCLUSION

This paper presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

## **5. REFERENCES**

- [1] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 9–18.
- [2] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "Gps estimation for places of interest from social users' uploaded photos," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2058–2071, 2013.
- [3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," IEEE Transactions on Multimedia, vol. 17, no. 6, pp. 907–918, 2015.
- [4] J. Sang, T. Mei, and C. Sun, J.T.and Xu, "Probabilistic sequential pois recommendation via check-in data," in Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2012.



- [5] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web*, vol. 5, no. 1, p. 5, 2011.
- [6] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in *Proceedings of 29th International Conference on AAAI*. AAAI, 2015.
- [7] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. ACM, 2014, pp. 659–668.
- [8] H. Yin, C. Wang, N. Yu, and L. Zhang, "Trip mining and recommendation from geo-tagged photos," in *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2012, pp. 540–545.
- [9] Y. Gao, J. Tang, R. Hong, Q. Dai, T. Chua, and R. Jain, "W2go: a travel guidance system by automatic landmark ranking," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 123–132.
- [10] X. Qian, Y. Zhao, and J. Han, "Image location estimation by salient region matching," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4348–4358, 2015.
- [11] H. Kori, S. Hattori, T. Tezuka, and K. Tanaka, "Automatic generation of multimedia tour guide from local blogs," *Advances in Multimedia Modeling*, pp. 690–699, 2006.