Literature Review

1. Master Thesis

- Goal: use machine learning algorithm to predict the gender based only on the text of the comment of the user
- Methods:
  - word-level bag-of-words approach (tf-idf, TF-IDFVectorizer)
    - benefits: straightforward approach of data representation → provides explanations

        - drawbacks: loss of semantics, large size of resulting matrix
        - algorithms: logistic regression, gradient boosting trees
    - word-level neural networks (word vectors using GloVe--pre-trained)
        - benefits: preserve the sequence of words
        - drawbacks: out-of-vocabulary words
        - algorithms: RNN (LSTM--long short-term memory architecture)
    - character-level neural networks (CNN)
        - benefits: preprocessing reduces memory and computational resources; handling out-of-vocabulary words
        - drawbacks: no concept of a word, large sizes of the neural network to increase the capacity to learn all information

```
![image-20181107211950875](/Users/xiaoqi/Library/Application Support/typora-user-images/image-20181107211950875.png)
```

The best machine learning classifier that was based on character-level convolutional network achieved an F1 score of 82.33% on a test set of 7.4 million comments from 6.9 thousand users.

2. Demographic Prediction Based on User's Browsing Behavior (2007)

- Dataset: Webpage click-though log data
- Related Words (based on authorship) --> little work on browsing history
    - linguistics writing and speaking styles
    - function words and parts-of-speech n-gram for author gender prediction --> 80% accuracy
- Goal: Given the webpage click-through log of some user with known demographic attributes, the problem is to find a general method to predict some users with unknown demographic attributes given their web-page click-through data
- Methods for (Users' Demographic Prediction)

1. Predict Webpages' Demographic tendency (content-based and categorial-based features)
2. Based on gender tendency of the webpages -- > use Bayesian framework
3. Smoothing approach (LSI to get similar webpages and users) --> indicate: similar user with similar browsing history --> similar gender tendency
    - Baseline algorithm: linear SVM
- Performance
    - Macro F1: 0.797 on gender prediction (30.4% improvements)

3. Your Cart tells you: Inferring Demographic Attributes from Purchase Data (2015)

- Dataset: purchase data (POS terminals)
- Proposal: propose a novel Structured Neural Embedding (SNE) model to learn representations from users' purchase data for predicting multiple demographic attributes simultaneously
- Motivation: Obtaining users' demographic attributes is crucial for retailers to conduct market basket analysis, adjusting marketing strategy, and provide personalized recommendations (However, most users are reluctant to provide detailed information or even refuse to register their demographics due to privacy and other reasons)
- Related Work
    - Predictable from different Behavioral data: web browsing, social media, mobile data...
    - followers on Twitter (likes on Facebook)
    - seldom practice on purchase behaviors in retail scenario
- Methods
    - Structured Neural Embedding (SNE) --> leverage the potential correlation between different attributes
        - characterize user with "bag-of-item" representations
        - feed this representation to a log-bilinear model

- Baseline algorithm:
  - SVD-single; SVD-structured
  - JNE (joint neural embedding) --> predict multiple attributes in parallel
- Performance: Finally, by learning the representations to predict multiple tasks in a structured way, our SNE can achieve the best performance in terms of all the evaluation measures under different observed label ratios. The improvement of SNE over the second best method (JNE) is significant (p value<0.01) in terms of all the evaluation metrics. (F1: 0.543 with 10% attributes missing)

4. You are where you go: inferring Demographics Attributes from Location Check-ins (2015)

- Purpose: (besides online behavioral data, such as "likes" on facebook, friendship relationship, and linguistic characteristics of tweets), in this paper, we investigate the predictive power of location check-ins for inferring users' demographics
- Proposal: propose a simple general location to profile (L2P) framework
  - extract fish semantics of users' check-ins --> spatiality, temporality and location knowledge
- Motivation:
  - commerce: the profile contributes significantly to link prediction, item recommendation and targeted advertising, which are crucial for most companies
  - uses: the profile is directive for content sharing, membership attachment, and trust establishment, which are pervasive for various personalized services
- Methods:
  - STC -- Spatiality, Temporality, and category-based method, which adopts spatial, temporal and category features in the tensor for prediction
  - STL -- comprehensively consider spatiality, temporality, and location knowledge

- Baseline: POI method (point of interest)
- Performance: STC -- 0.77 F1 score; STL -- 0.81 F1 score
- Tensor Factorization: our work is the first to apply tensor factorization for profile inference in real world online social networks where features are organized in a three-way tensor, which consists of user, context (spatiality and temporality) and location knowledge.

5. Predicting website audience demographics based on browsing history

- Background:
    - "personalized advertising" is used to denote the practice of showing individual web users advertising messages relevant for them as opposed to randomly selected messages
    - obtain the demographic distribution of a website as a whole. This is crucial for online publishers because they sell advertising space on their websites and advertisers with to know what kind of audience their ads will be exposed to
- Algorithm Type:
    - Contextual: (contextual techniques aim to predict demographic distribution of the entire website and cannot detection demographic difference among individual website visitors)
        - set of words on the webpage; words in title and sectioning
        - Latent Dirichlet Allocation for information retrieval, logistic regression
    - Behavioral:
        - search terms
        - syntactic and semantic information from urls
        - content based and category based from webpage
        - web search queries
        - clickstream data
        - clickstream patterns including set of websites visited, day of the

        week and time of the visits, intensity and frequency
        - browsing history including social networks
- Conclusion:
    - success of the previous studies predicting online audience demographics was affected by the algorithm used, input features, the demographic variable being predicted and the number of classes predicted.
    - URLs of visited websites, the day of the week and the time of the visit, the total number of websites visited and the number of online ads clicked are helpful for predicting demographics of online audiences

6. Inferring the Demographics of Search Users

- Background: personalizing web search results or related services such as query suggestion and query completion. We take this line of previous work to the next level by showing that the demographics of users can be automatically predicted based on their past queries.
- Motivation: Today it is quite common for web page content to include an advertisement. Since advertisers often want to target their message to people with certain demographic attributes, the anonymity of Internet users poses a special problem for them.
- Related work:
    - for instance, females on average submit long queries.
    - male and female interact differently with sponsored search result
    - reranking the search results based on users' genders may enhance their experience in particular for ambiguous queries
- Method:
    - both Facebook Likes and search queries can be translated into a common representation via mapping to ODP(Open Directory Project) categories -- can be seen as coarse grained representation
    - Facebook Likes need to be matched against queries. We achieve that by developing a common representation for Facebook Likes and search

queries within the Open Directory Project (ODP) categories
- Conclusions: Demographics of search users can be accurately predicted based on models trained on an independent data, achieved 80% AUC for gender

7. Gender Prediction Using Browsing History

- Related Work:
  - textual data: reviews, blogs, comments, tweets, emails --> writing style or speaking style
  - browsing behaviors: pages or multimedia content user visited; user comments on videos and the relationship of the people watching the same video
- Features:
  - category-based features (tf-idf)
  - topic-based features (LDA--Latent Dirichlet allocation)
  - Time features
  - Sequential features (order of viewing pages--hypo: men tend to change between categories more frequently)
  - combining features
- Evaluation Metrics: F1 score $F1 = \frac{2pr}{p+r}$. We judged the performance of the proposed and a baseline method in terms of precision p, recall r, and F1 score. For each class (male, female), precision is defined as the number of correctly predicted cases divided by the number of all predictions of this class. Recall is defined as the number of correctly predicted cases divided by the number of all cases of this class. F1 is the harmonic mean of precision and recall
- Performance:
  - Baseline: linear SVM -- F1: 0.695
  - Topic-based + Time -- F1: 0.75
  - Topic-based + Sequential -- F1: 0.735
  - Topic-based + Sequential + Time (all) -- F1: 0.805

8. Classifying Latent User Attributes in Twitter

- Motivation: important applications in advertising, personalization, and recommendation
- Finding: status message content is more valuable in inferring latent author attributes
- Features: simply by content and behavior of their postings (content and style of their writing)
- Methods: Stacked (sociolinguistic-feature models + Ngram-feature) --> accuracy:72.33
    - Sociolinguistic-based features for GENDER expressed as relative frequency of females and males.

Feature	*#female/#male*
Emoticons	3.5
Elipses	1.5
Character repetition	1.4
Repeated exclamation	2.0
Puzzled punctuation	1.8
OMG	4.0

9. Author Gender Prediction in an Email Stream using Neural Networks

- Abstract: Experiments show that it is effectively able to discriminate gender using both stylometric and word count features, with the word count features providing superior results

- Related works: multiple linguistic features have been determined, such as character usage, writing syntax, functional words, and word frequency
    - Women tend to use more emotionally charged language as well as more adjectives and adverbs, and apologize more frequently than men. Men use more references to quantity and commit a greater number of grammatical errors.
    - Business-related emails have less gender-preferential language than blogs, making business emails harder to classify than blogs and lowering the expected accuracy
- Method: Neural Network
- Results: When the best parameters were found, the stylometric features achieved 88% accuracy and the word based features achieved around 95% in comparison to approximately 56% accuracy for the traditional Balanced Winnow using both feature sets

10. Customer gender prediction based on E-commerce data

- Abstract: features based on their catalog viewing data on e-commerce systems, such as the data and time of access, list of categories and products viewed
    - 81.2% on balanced accuracy and 81.4% on macro F1
- Features:
    - basic: viewing time, products/categories features,
    - advanced: products/categories sequence and transfer features
- Motivation: Personalization is mainly based on two types of data:
    - historical data (e.g. previous item viewed or purchased)
    - demographic-based methods
- Related work:
    - authorship studies: writing style using various types of features, such as lexical, syntactic, or content-based features
    - Recently, due to the growth of Internet and online communication channels, the focus has been moved to computer mediated

    - communication contents, such as email, blogs, comments
    - browsing behaviors; the content and hyperlinked structure
- Results: Ramdom fortets: 81.4; SVM: 78.8; BayesNet 78.6. (Macro F1)

11. Inferring Gender from Names on the Web: A comparative evaluation of gender detection methods

- Abstract: Our findings show the the performance of name-based gender detection approaches can be biased towards countries of origin and such biases can be reduced by combining name-based an image-based gender detection methods
- Methods: we propose mixed methods that combine name-based detection methods with an image-based face recognition approach.
- Results: the error rates strongly depend on the country of residence of an individual.

Table 2: Accuracy of various gender detection methods for people from different countries. For most countries mixed approaches perform best.

	# instances	SSA	IPUMS	Sexmachine	Genderize	Face++	Mixed1	Mixed2
United States	419	0.82	0.76	0.84	0.83	0.91	0.91	0.90
China	113	0.20	0.11	0.67	0.28	0.65	0.50	0.56
United Kingdom	96	0.94	0.92	0.92	0.94	0.81	0.98	0.94
Germany	82	0.87	0.88	0.96	0.94	0.87	0.96	0.93
Italy	75	0.93	0.92	0.94	0.98	0.79	0.99	1
Canada	60	0.87	0.77	0.86	0.91	0.90	0.96	0.93
France	58	0.93	0.92	0.80	0.96	0.81	0.97	1
Japan	56	0.79	0.70	1	0.90	0.62	0.91	0.94
Brazil	44	0.29	0.29	0.15	0.44	0.81	0.90	0.93
Spain	39	0.96	0.92	0.92	1	0.92	1	1
Australia	31	0.89	0.89	0.90	0.86	0.86	0.94	0.93
India	29	0.67	0.17	0.71	0.78	0.83	0.83	0.93
South Korea	27	0.04	0.00	0.58	0.11	0.74	0.37	0.66
Switzerland	25	0.78	0.70	0.56	0.83	0.88	0.90	0.92
Turkey	21	0.43	0.14	0.79	0.81	0.86	1	1

12. Gender prediction based on semantic analysis of social media images

- Abstract:
    - distribution of semantics in the picture coming from the whole feed of a person to estimate gender
    - the gender signal can indeed be extracted from the users image feed (75.6% accuracy)
- Related work
    - estimate gender from textual analysis of diverse sources such as tweets, hashtags, psycho-linguistic features, conceptual attributes; first name analysis
    - profile picture face analysis alone is not sufficient for fully reliable gender estimation
- Semantics
    - claim that even in case where a user's face is not portrayed in his/her profile picture, the choice of subject for such picture is correlated with the user's gender
    - 25 categories: adult, animal, baby, beach...
    - male seem to be cat lovers, female users seem to prefer Dog. male post more vehicles(car and motorcycle) while female have more profile pictures with friends (two people) and landscapes, both rural and urban
- Conclusion: provide a strong gender prediction cue (75.6% accuracy), which proved to be complementary to traditional textual analytics (88% accuracy visual+textual features)

13. Improving Gender Prediction of Social Media via Weighted Annotator Rationales

- Introduction: prediction user characteristics can help answer important social science questions and support many commercial applications including targeted computational advertising to math user interest profile from Twitter or Facebook, detecting fraudulent product reviews or branding analytics.

(a) Male rationales

(b) Female rationales

Figure 2: Gender rationales (word ngrams of size ≤ 3 and frequency ≥ 2). The size of each rationale reflects the frequency of being highlighted.

Approach	Users	Tweets	Features	Accuracy
Rao et al. [19]	1K	0.4M (4K)	BoW, socio-ling, combined	**0.69** 0.71 0.72
Burger et al. [5]	184K	4M (22)	user names, char ngrams	0.92 **0.75**
Bergsma et al. [4]	400 1M	4B (500)	bootstrapped bootstacked	**0.72** 0.87
This work	357	70K (200)	T (only) $T' + R$	**0.66** **0.75**

Table 2: The overview of the existing approaches for gender classification in social media.

- Conclusion: We proposed several readily-replicable new models for gender classification of social media users for English and Spanish that outperform the state-of-the-art models learned exclusively from user data.

14. Gender Inference of Twitter Users in Non-English Contexts

- Gender in Japanese, in contrast, could not be reliably inferred with any reasonable accuracy (on average) despite numerous attempts to preprocess the tweets and tune the classifier to accommodate the language's complex orthography. This indicates that existing approaches may not generalize well to language systems with thousands of distinct unigrams (as opposed to tens or hundreds in the other languages considered)
- The features we employed were: k-top words, k-top digrams and trigrams, k-top hashtags, k-top mentions, tweet/retweet/hashtag/link/mention frequencies, and out/in-neighborhood size.

Table 2: The accuracy of the SVM-based classifier on each of the language datasets. In the case of Japanese, the performance is given for both the tokenized and un-tokenized versions of the dataset. (Note that tokenization did not affect overall accuracy.)

Language	Male	Female	Overall
French	0.79	0.73	0.76
Indonesian	0.87	0.80	0.83
Turkish	0.89	0.85	0.87
Japanese (t)	0.50	0.76	0.63
Japanese (u)	0.58	0.68	0.63

15. The eyes of the beholder: Gender prediction using images posted in Online social networks

- Abstract: Abstract—Identifying user attributes from their social media activities has been an active research topic. The ability to predict user attributes such as age, gender, and interests from their social media activities is essential for personalization and recommender systems. Most of the techniques proposed for this purpose utilize the textual content created by a user, while multimedia content has gained popularity in social networks. In this paper, we propose a novel algorithm to infer a user's gender by using the images posted by the user on different social networks.
- Introduction:
  - OSNs such as Instagram and Pinterest that are majorly image based have gained popularity with almost 20 billion photos already been shared on Instagram and an average of 60 million photos shared daily.
  - Intuitive findings: for male users, they are more interested in electronics, buildings, mens clothes and so on. Female users are more likely to post

inboards that are related to jewelry, women clothes, gardening and so on.

- Results: evaluate the performance by combining both user posting behavior and visual content based features -- achieve better overall performance -- achieved F-measure of around 72% using both types of features

Table VI: Performance of using both posting behavior and posted content.

Class	Accuracy	Precision	Recall	F-Measure
Female	0.688	0.733	0.688	0.71
Male	0.75	0.706	0.75	0.727
Avg	0.719	0.72	0.719	0.718

16. Minimalistic CNN-based ensemble model for gender prediction from face images

Table 1. Gender recognition results in an uncontrolled environment.

Authors	Test dataset	Method	Cross-Dataset	Accuracy
Shan (2012)	LFW	LPB + AdaBoost	No	94.81%
Shih (2013)	color FERET + LFW	AAM + Bayesian	No	86.50%
Tapia and Perez (2013)	LFW	multiscale LBP + SVM	No	98.01%
			Yes	95.60%
Bekios-Calfa et al. (2014)	LFW	appearance-based + LDA	Yes	79.11%
Levi and Hassner (2015)	Adience	CNN	No	86.80%
Jia and Cristianini (2015)	LFW	multiscale LBP + C-Pegasos	Yes	96.86%

- Conclusion: designed a CNN-based ensemble model for gender recognition from face images
    - the record performance of 97.31% on LFW dataset has been set using the ensemble of 3 CNNs