Identifying Gender From SMS Text Messages

Pavithra C P
PKD17CSCL10

Guided By
Shibily Joseph
Asst. Professor
Department of Computer Science and Engineering
GOVERNMENT ENGINEERING COLLEGE, PALAKKAD

April 11, 2017

Outline

- Introduction
- Problem Formulation
- System Architecture
- System Implementation
- Comparison
- Result
- Conclusion
- References

Introduction

- Short message service (SMS) has become a very popular medium for communication due to its convenience and low cost.
- As of December 2012, U.S. wireless users sent and received an average of 6 billion text messages per day.
- This growth also encourages various kinds of misuses.
- Online communities are vulnerable to deceptive attacks, receiving false information, etc.

Introduction

- It easy to provide a false name, age, gender, and location in order to hide ones true identity.
- It becomes imperative to design an efficient method for identity tracing in cyberspace forensics.
- Gender identification is on the focus here.

Introduction

Example:

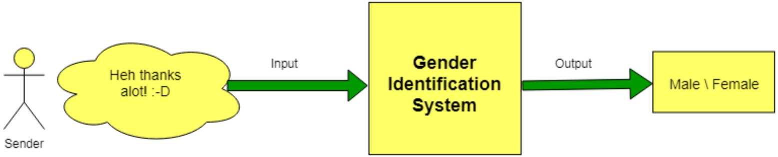

Figure: Input-Output Flow

Problem Formulation

- A question address in text based Internet forensics is the following: *given a short text document, can we identify if the author is a man or a woman?*

Problem definition:

Given a short text message , identify the gender of the sender using machine learning techniques.

The problem is a binary classification problem involving text based features.

Overview of Work

- Comparison of two approach for gender identification is performed:
  - Based on n-gram feature
  - Based On Manually extracted 5-set of features
- Comparison on two classification algorithm:
  - Naive Bayes
  - Support Vector Machine(SVM)
- Identify which feature is important to predict the gender.
- Testing and Prediction for one input is performed.
- Implementation Language : Python3

System Architecture

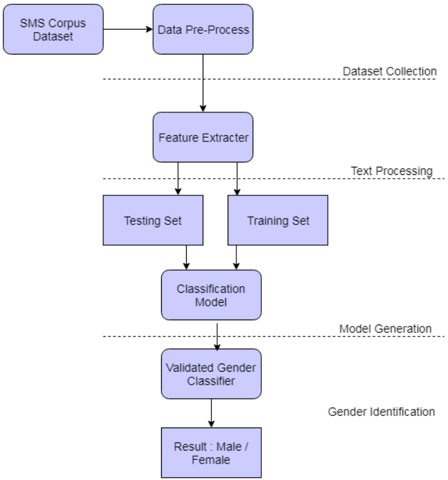

Figure: Gender Identificaion Process

System Implementation

Mainly Consist of 4 Modules:

- Data Collection
- Text Processing
- Model Generation
- Gender Identificaion

Data Collection

- NUS SMS Corpus is used as dataset, which contains 55,585 SMS text messages.
- The final dataset consist of 28,288 male author messages and 12,647 female author messages.
- In order to clean up the data the gender unknown messages, VCARD messages and chinese messages were removed.
- Total 40935 messages are there in final dataset.
- The first module is same for both the approaches.

Cont..

"@time": "2003/4"}}, {"@id": 15527, "text": {"$": "Cable can jus buy one
rite?"}, "source": {"srcNumber": {"$": 81}, "phoneModel": {"@manufactuer":
"unknown", "@smartphone": "unknown"}, "userProfile": {"userID": {"$": 81},
"age": {"$": "unknown"}, "gender": {"$": "unknown"}, "nativeSpeaker": {"$":
"unknown"}, "country": {"$": "SG"}, "city": {"$": "unknown"}, "experience":
{"$": "unknown"}, "frequency": {"$": "unknown"}, "inputMethod": {"$":
"unknown"}}}, "destination": {"@country": "unknown", "destNumber": {"$":
"unknown"}}, "messageProfile": {"@language": "en", "@time": "unknown", "@type":
"unknown"}, "collectionMethod": {"@collector": "howyijue", "@method": "unknown",
"@time": "2003/4"}}, {"@id": 15528, "text": {"$": "Did noe valentine is oso
friendship day.. eh heh, long live our friendship and A6!"}, "source":
{"srcNumber": {"$": 81}, "phoneModel": {"@manufactuer": "unknown",
"@smartphone": "unknown"}, "userProfile": {"userID": {"$": 81}, "age": {"$":
"unknown"}, "gender": {"$": "unknown"}, "nativeSpeaker": {"$": "unknown"},
"country": {"$": "SG"}, "city": {"$": "unknown"}, "experience": {"$":
"unknown"}, "frequency": {"$": "unknown"}, "inputMethod": {"$": "unknown"}}},
"destination": {"@country": "unknown", "destNumber": {"$": "unknown"}},
"messageProfile": {"@language": "en", "@time": "unknown", "@type": "unknown"},
"collectionMethod": {"@collector": "howyijue", "@method": "unknown", "@time":
"2003/4"}}, {"@id": 15529, "text": {"$": "Thru icq. :)"}, "source":

Figure: Dataset

Text Processing

- In second module, text processing and feature extraction is performed.
- Cleaned dataset is the input and the extracted features set is the output to these module.
- In First approach five sets of gender-related features are considered:
    - Character based
    - Word based
    - Syntactic
    - Structure
    - Function words

First Approach

- Manually extracted 5-set of features.
- Total : 59 Features are extracted
  - Character based : 6
  - Word based : 4
  - Syntactic : 10
  - Structure : 4
  - Function words: 35
- Saved as a .csv file for testing and training.

Cont..

gender	charcount	letter	upper	spcl	digts	space	wrds	long	short	avg lngth	qutn	commas
0	18	15	1	15	0	2	3	2	1	5	0	0
0	3	2	1	2	0	0	1	0	1	3	0	0
0	13	10	1	10	0	2	3	0	2	3	0	0
0	16	11	1	11	0	2	3	0	0	4	0	0
0	14	9	1	9	0	1	2	1	0	6	0	0
0	27	19	1	19	0	4	5	2	2	4	1	0
0	19	15	1	15	0	2	3	2	1	5	0	0
0	12	11	1	11	0	1	2	1	1	5	0	0
0	16	13	1	13	0	2	3	1	1	4	0	0
0	8	7	1	7	0	1	2	0	1	3	0	0
0	14	12	1	12	0	2	3	0	1	4	0	0
0	14	12	1	12	0	2	3	0	1	4	0	0
0	9	8	1	8	0	1	2	0	1	4	0	0
0	14	12	1	12	0	2	3	0	1	4	0	0
0	11	10	1	10	0	1	2	1	0	5	0	0
0	5	4	1	4	0	0	1	0	0	5	0	0
0	16	13	1	13	0	2	3	2	1	4	0	0
0	16	13	1	13	0	2	3	2	1	4	0	0
0	10	9	1	9	0	1	2	1	1	4	0	0

Figure: Feature set

Second Approach

- To Implement n-gram model, Tf-Idf vectorization method is used.
- From sklearn feature extraction module TfidfVectorizer is imported.
- In which, n-gram range is set as 1 and 2. ie, unigram and bigram is considered.
- (eg)

  vect = TfidfVectorizer(ngram_range=(1,2),max_features=100)

Cont..

	about	ah	all	also	am	and	are	are you	ask	\
0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
1	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
2	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
3	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
4	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
5	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
6	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
7	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
8	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
9	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
10	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
11	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
12	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
13	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
14	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
15	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
16	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
17	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
18	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
19	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
20	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
21	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
22	0.000000	0.0	0.0	0.0	0.0	0.791973	0.000000	0.000000	0.0	
23	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
24	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
25	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
26	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
27	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
28	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
29	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	
...	...	...	...	...	...	...	...	...	...	

Figure: Feature set using n-gram

Model Generation

- In third stage classification model is created.
- Feature set is the input to these stage and output is a binary classifier.
- Two classification algorithms are used.
  - Naive Bayes
  - Support Vector Machine
- Feature Set is split for training and testing in 1/3 ratio.
- Same for both approaches and the classification model is created.

Gender Identification

- Test for both approaches and Performance is measured using:
  - Accuracy
  - Precision
  - Recall
  - F-measure
- Prediction is also performed for one input text message.

Comparison

Model	Algorithm	Accuracy	Precision	Recall	F-Measure
N-Gram Model	SVM	71.07	0.98	0.71	0.82
	NB	64.52	0.69	0.77	0.73
Feature Extracted Model	SVM	77.75	0.96	0.77	0.85
	NB	68.11	0.72	0.79	0.75

Figure: Comparison of 2-approaches

Best Feature

Features Included	No: of Features	Accuracy
Character Based	6	71.74 %
Word Based	4	69.69%
Syntactic Based	10	76.48%
Structural Based	4	68.88%
Functional Based	35	71.75%

Figure: Comparison- only one feature

Combination of Features	No.of Features	Accuracy
Character, Word	10	71.69%
Character, Syntactic	16	76.59%
Character, Structural	10	72.23%
Character, Functional	41	72.67%
Word, Syntactic	14	75.82%
Word, Structural	8	70.22%
Word, Functional	39	72.02%
Syntactic, Structural	14	77.02%
Syntactic, Functional	45	76.66%
Structural, Functional	39	72.74%

Figure: Comparison- two feature set

Cont..

Combination of Features	No: of Features	Accuracy
Character, word, Syntactic	20	76.31%
Character , Word, Structural	14	72.31%
Character , Word, Functional	45	72.73%
Character , Syntactic, Structure	20	76.92%
Character , Syntactic, Functional	51	76.86%
Character, Structural, Functional	45	73.05%
Word, Syntactic, Structure	18	76.24%
Word, Syntactic, Functional	49	76.91%
Word, Structural, Functional	43	73.07%
Syntactic, Structure, Functional	49	77.36%

Figure: Comparison- three feature set

Cont..

Combination of Features	No: of Features	Accuracy	Not Included
Character, word, Syntactic, Structural	24	76.74	Functional
Character , Word, Syntactic, Functional	55	76.81	Structural
Character , Word, Structure, Functional	49	73.23	Syntactic
Character , Syntactic, Structure, Functional	55	77.64	Word
Word, Syntactic, Structure, Functional	53	77.64	Character

Figure: Comparison- four feature set

Result

- Feature extracted model is better than n-gram model.
- In both approaches, support vector machine performs more than naive bayes classification algorithm.
- Syntactic Feature is important among 5-set of features.

Conclusion

- Gender identification is one of the technique to identify the true identity of a person.
- In these Work, mainly focusing two approaches:
  - Feature extracted method.
  - N-gram method.
- Two classification algorithms are used.
- From the study, its conclude that Feature extraction method with SVM algorithm gives better result for gender identification.

Reference

- [1] Shannon Silessi, Cihan Varol et.al "Identifying Gender From SMS Text Messages" *2016 15th IEEE International Conference on Machine Learning* Sam Houston State University Huntsville, TX, USA.
- [2] N. Cheng et al. Author gender identification from text, *Digital Investigation* Volume 8, Issue 1, pp. 78-88, Jule 2011.
- [3] S. Argamon et al. Automatically profiling the author of an anonymous text, *Communications of the ACM - Inspiring Women in Computing* Volume 52, Issue 2, pp. 119-123, February 2009.
- [4] J. Soler and L. Wanner. How to Use Less Features and Reach Better Performance in Author Gender Identification, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014, pp. 1315-1319.

Thank you!