A Robust Gender Inference Model for Online Social Networks and its Application to LinkedIn & Twitter

by Athanasios K. Kokkos, and Theodoros Tzouramanis

Abstract:

Online social networking services have come to dominate the dot com world: Countless online communities coexist on the so-named Social Web. Some typically characteristic user attributes, such as gender, age group, sexual orientation, are not automatically part of the profile information that is provided by default in some of these online services, and in some cases they can even be deliberately and maliciously falsified. This paper looks into how the automated inference of the gender of a user of online social networks can be achieved by means of the application to this user's written text data of a combination of natural language processing and classification techniques. The paper aims to show that the use of a support vector machine classifier and a part-of-speech tagger makes it possible to infer the gender of an online social network user and that this can be done through the classification of down to one single short message included in a profile, quite independently of whether this message does follow a structured and standardized format (as with the attribute summary in LinkedIn) or does not (as with the micro-blogging postings in Twitter). Extensive experimentation on LinkedIn and Twitter has indicated a very marked degree of accuracy of the proposed gender identification scheme of up to 98.4%. This represents a higher level of accuracy than had been achieved by research in the field. The approach proposed herewith may lend itself to applications that could prove significant in a number of fields such as the advertising world; personalization and recommendation; security, forensics and cyber investigation; etc. To the best of the authors' knowledge this paper is actually the first to address the gender identification problem in online social networks that are used for business-only interactions, such as LinkedIn.

Contents

1. Introduction

In the age of the omnipresent social web, personal information is exchanged, shared and transformed in a multitude of ways, as the amount of time which hundreds of millions of people spend being interconnected to various online communities and social networks on a daily basis is ever increasing: in this process, they reveal all manners of details about themselves to the point of exposing their innermost thoughts and feelings, private personal information, preferences, beliefs and concerns. Hence the significant amount of research interest in aspects of information retrieval and natural language processing, aiming at extracting the information relevant to their purpose from the source of knowledge which this massive amount of online data represents.

This study focuses on ways of achieving, in the context of online social networking platforms and through a classification of available users' written texts, the inference of users' personal attributes that

have been kept private or hidden. More precisely, the proposed model deploys psycholinguistic tools in an endeavour to determine the gender of the user in cases when that item of information might have been kept hidden or might have been deliberately falsified. The problem comes under the wider umbrella of authorship profiling (Argamon *et al.*, 2009): that is the case of the automated prediction of users' latent demographic features achieved by applying advanced text mining techniques. The proposed model uses only a small features-set to classify text data, which has been proved to serve adequately for the purpose of gender inference through experimentation with a large trial group of users.

The main advantage of the proposed scheme is that it performs efficiently even in the case of online social networks that are used for business interaction – in which the users create formal professional profiles and where language is more standardized and neutral – and that it is not necessarily restricted to those networks which are limited to personal and social communication purposes. To the best of the authors' knowledge this study is the first to address the issue of attribute disclosure and automated user profiling in LinkedIn; the aim of the study is to demonstrate that this is achievable even with limited resources. For a performance comparison of the proposed scheme with previous research carried out in this domain, an implementation of the prototype of the proposed model is also applied to the popular platform of Twitter, for which this work reports higher levels of accuracy than were achieved by most of the previous research efforts in this field of investigation. Another advantage of the proposed model is that it can be easily modified to serve applications pertaining to any form of written text because it makes use of content-based sociolinguistic features to a much greater extent than it does of n-gram features. The concept proposed in this paper opens avenues to possibilities of extending the width of its application towards making it possible to infer further undisclosed personal attributes. It may also be applied to other online social networks.

The proposed methodology can be adopted in diversified application domains, from security tools for intelligence services that might target authorship profiling to the fighting of cybercrime, to data mining technologies of important value in commercial terms to advertisers and decision making environments.

The new scheme could further prove useful as part of a larger user identity authentication service consisting of numerous sophisticated and complex data mining components that would serve social network providers as an efficient protection mechanism (Chaski, 2005) for their members, as with the prevention of bullying, harassment and abusing the trust of the vulnerable members of society.

The article unfolds as follows. Section 2 develops around a new methodology for the inference of the gender of an online social network by elaborating an advanced data analysis technique from the users' written text data. Section 3 describes the implementation of a prototype text mining system to infer the gender of a user on LinkedIn and Twitter on the basis of the proposed methodology. Section 4 discusses issues related to the evaluation of the proposed scheme. Section 5 analyses the advantages and limitations of the related work and argues for the efficiency of the new proposal by comparing its accuracy to the accuracy of earlier research in the field. Finally, Section 6 summarizes the article and suggests some directions for further research.

2. The Proposed Methodology

2.1 Theoretical background

The paper puts forward a new approach for the inference of the gender of the users of online social networks. Since numerous online social networks do not provide information about the gender of their members (with LinkedIn and Twitter being two of the most popular such examples), building text mining techniques to uncover this type of undisclosed attributes on the basis of the user-related information which is available online represents a challenge. In other words, the challenging task consists of the construction of a machine learning algorithm that will be trained to follow a gender classification model from a set of known written text samples produced by authors on either side of the gender divide, and then to use this algorithm in order to infer the gender of any of the users of these social networks.

The proposed methodology relies on several studies in cognitive psychology, computational linguistics and computer forensics that argue that human beings display unique, or near-unique behavioural patterns and that psycholinguistic analysis techniques, in the sense of analyzing the psycho-linguistic properties of texts, and machine learning techniques (i.e. machine learning classifiers that learn to classify text data) can be combined to infer user gender, age group, first language, educational background, religious affiliation, possible sexual orientation and, generally, to reveal many social and psychological aspects of an individual's profile by analyzing the text s/he produces in her/his daily written interaction.

Since the present paper's focus is on gender inference, the authors find it necessary to provide a little clarification with regard to the context in which they base notions of gender along the aforementioned women-men binary boundary [1] by briefly elaborating on the background against which perceptions of behaviour as determined by gender stand in broad terms. Research in the field of human psychology in Broveman *et al.*, 1972; Crawford 1995; Eagly 1987; Eagly and Steffen 1984; and Greenwald and Banaji 1995 indicate that women and men adopt different and almost unique gender-based behavioural patterns in their oral and written interactions. In addition, a number of studies discuss the emotional content and the tone of a given written text and show that certain words can function as markers of emotional, psychological, and cognitive states; these may therefore be chosen to indicate substantial gender differential features. The studies Eagly and Steffen, 1984 and Greenwald and Banaji, 1995 have examined and compared feminine and masculine behavioural and attitudinal dimensions and have subsequently drawn borders along the line of two distinct gender stereotypes. This literature tends to indicate that, by comparison to women, men's language use patterns that are characterized by more marked expressions of independence and assertions of vertically hierarchical power and include more strongly assertive, aggressive, self-promoting features, rhetorical questions, authoritative orientation and challenges. In contrast, in comparison to men, women tend to express themselves with the use of more emotional language, the use of more frequently emotionally intensive adverbs and affective adjectives

(such as really, quite, adorable, charming, lovely), and also that their language expresses much more attenuated assertions, apologies, questions, personal orientation and support. The studies Broverman *et al.*, 1972, and Crawford, 1995 have also argued that men, in their use of language, come over as being more proactive, by directing speech at solving problems, while women come over as being more reactive to the contributions of others, agreeing, understanding and supporting. On the other hand, women are deemed more expressive of certain emotions and more concerned about maintaining intimacy in their close relationships, while men are found to be better at controlling their nonverbal expressions, and to be more concerned with maintaining autonomy in their relationships.

Women's and men's propensity to behave along gender-typical patterns and the reason why their behaviour tends to confirm the gender stereotypes comes down to the tendency to act in accordance with the social roles which they have been allocated (Eagly, 1987), which are conditioned and differentiated across gender boundaries. These social roles of men and women relate to different expectations and require different skills. Each gender adopts different behavioural and expressive patterns in interaction, mainly based on the way society moulds the individuals from their birth. Therefore, the requirements of the social roles that women and men enact on a daily basis encapsulate the key idea to perceiving the differences in gender behaviour. According to Rubin and Greene, 1992, these significant differences between women and men are also present in their written communication patterns. Therefore, if it is possible to integrate measures of emotional and psychological states, it will improve significantly the robustness of the proposed model for gender identification through written text.

Additionally, from the computational linguistics and computer forensics perspective, several studies, as for example Argamon *et al.*, 2009, employ stylometric measures (i.e., style markers) for discriminating authors, and already over 1,000 such features have been proposed for this task, which include: character-based, word-based, function words, syntactic-based and structural-based stylometric features. Research in the field of authorship profiling and attribute disclosure (Diederich *et al.*, 2000) also shows that the Support Vector Machine (SVM, Joachims, 1998) is the most suitable machine learning classifier

for binary text classification problems, such as gender prediction, since in most cases it outperforms conventional classifiers and produces the most robust data classification results.

The strategy that will be followed in the proposed approach will exploit both *content-based features* (e.g., words related to specific feeling classes that can be markers of emotional, psychological and cognitive internal states of a person at a given time), and *traditional-style features* (e.g., markers of female and male writing styles, such as character-based features, word-based features, syntactic features and function words) that are strong gender indicator cues, in order to establish a robust gender discrimination tool that will provide accurate inference results.

In the case of LinkedIn, content-based psycho-linguistic and traditional-style gender-preferential features will be extracted from the user profile's attribute *Summary* – that is a textual field in which the LinkedIn members describe their professional profile (LinkedIn, 2014)– to predict the user's gender, while in the case of Twitter the same kind of features will be extracted from the user's tweets. Each text sample will be transformed into a multidimensional feature vector, with each feature contributing to classifying the author of the text under the corresponding gender category. Then a machine learning text classifier will be constructed that will be trained by selected gender pre-classified data in order to predict the users' gender in the online social network platforms concerned.

2.2 Features-set selection

The literature on pattern-recognition emphasizes that the most critical step that will be taken on any pattern-matching task is probably that of finding the best suited features-set to input into the right classification procedure. On this basis, the recent studies Comey *et al.*, 2002 and Cheng *et al.*, 2011 recognize that the word-based features and function words are the linguistic features that can play the most important role in the gender inference process. These linguistic features include also content-based features which can achieve efficient text classification results, as they are able to successfully distinguish

between the genders. In the present study, a combination of several types of features was selected to build a mixed set of 423 style marker features, which include character-based features, word-based features, psycho-linguistic features, syntactic features, gender-preferential language features, and finally function words. A quick view of this 423-features-set is demonstrated in Table 1.

Character-based features (4)
1. Total number of characters (C)
2. Total number of letters (a-z) / C
3. Total number of upper characters / C
4. Total number of white-space characters / C
Word-based features (21)
5. Total number of words (N)
6. Average length per word (in characters)
7. Vocabulary Richness (total different words/ N)
8. The number of net abbreviation / N
9. Words longer than 6 characters / N
10-25. psycho-linguistic features (16 measures that indicate the emotional state of people). More details on Table 2.
Syntactic features (4)
26. Number of question marks (?) / C
27. Number of multiple question marks (???) / C
28. Number of exclamation marks (!) / C
29. Number of multiple exclamation marks (!!!) / C

Gender-preferential language features (9)
30. Number of words ending in *able* / N
31. Number of words ending in *al* / N
32. Number of words ending in *ful* / N
33. Number of words ending in *ible* / N
34. Number of words ending in *ic* / N
35. Number of words ending in *ive* / N
36. Number of words ending in *less* / N
37. Number of words ending in *ly* / N
38. Number of words ending in *ous* / N
Function words-based features (385)
39-41. Number of articles / N
42-117. .Number of pronouns / N
118-164. Number of auxiliary verbs / N
165-187. Number of conjunctions / N
188-299. Number of interjections / N
300-423. Number of adverbs and prepositions/ N

Table 1: A more detailed view of the selected 423 features-set.

On a more detailed level, the character-based features subset consists of 4 features: the total number of characters (C); the total number of lower-case characters / C; the total number of upper characters / C; and the total number of white-space characters / C. From the wide range of the word-based features category that have been proposed in the literature, 5 statistical metrics (Lines 5 to 9 in Table 1) were

chosen and 16 psycho-linguistic features extracted from PsychPage, 2014 (Line 10-25 in Table 1). The psycholinguistic features include on the one hand words related to pleasant emotional states (e.g. open, happy, alive, good, love, interested, positive, strong, etc.) and include on the other hand words related to difficult or unpleasant emotional states (e.g. angry, depressed, confused, helpless, indifferent, afraid, hurt, sad, etc.). These features are usually significant indicators of emotional states, thus reliable pointers to people's gender, since, it will be recalled from Subsection 2.1, communication patterns across the gender divide are marked by the way either side enact their perceived social roles and consequently result accordingly in the manifestation of differences in the way feelings find their expression. The manifestations of this dynamic can therefore be used for the purpose of classification (Argamon *et al.,* 2009; Cheng *et al.,* 2011; Mauss and Robinson, 2009; and Shields *et al.,* 2006). The 16 psycho-linguistic features are listed in more detail in Table 2. As the Table shows, every one of the 16 features is related to a number of words that relate in turn to the same area of emotions. The value of every feature can be computed by measuring the frequency of use of the words that pertain to the corresponding emotional and psychological state.

Emotional & psychological state	**Words pertaining to feelings included into the 16 psycho-linguistic features**
Pleasant feelings	
Open	Understanding, confident, reliable, easy, amazed, sympathetic, interested, satisfied, receptive, accepting, kind
Happy	Great, gay, joyous, lucky, fortunate, delighted, overjoyed, gleeful, thankful, important, festive, ecstatic, satisfied, glad, cheerful, sunny, merry, elated, jubilant
Alive	Playful, courageous, energetic, liberated, optimistic, provocative, impulsive, free,

	frisky, animated, spirited, thrilled, wonderful
Good	Calm, Peaceful, at ease, comfortable, pleased, encouraged, clever, surprised, content, quit, certain, relaxed, serene, free and easy, bright, blessed, reassured
Love	Loving, considerate, affectionate, sensitive, tender, devoted, attracted, passionate, admiration, warm, touched, sympathy, close, loved, comforted, drawn toward
Interested	Concerned, affected, fascinated, intrigued, absorbed, inquisitive, nosy, snoopy, engrossed, curious
Positive	Eager, keen, earnest, intent, anxious, inspired, determined, excited, enthusiastic, bold, brave, daring, challenged, optimistic, re-enforced, confident, hopeful
Strong	Impulsive, free, sure, certain, rebellious, unique, dynamic, tenacious, hardy, secure
Unpleasant feelings	
Angry	Irritated, enraged, hostile, insulting, sore, annoyed, upset, hateful, unpleasant, offensive, bitter, aggressive, resentful, inflamed, provoked, incensed, infuriated, cross, worked up, boiling, fuming, indignant
Depressed	Lousy, disappointed, discouraged, ashamed, powerless, diminished, guilty, dissatisfied, miserable, detestable, repugnant, despicable, disgusting, abominable, terrible, in despair, sulky, a sense of loss
Confused	Upset, doubtful, uncertain, indecisive, perplexed, embarrassed, hesitant, shy, stupefied, disillusioned, unbelieving, sceptical, distrustful, misgiving, lost, unsure, uneasy, pessimistic, tense
Helpless	Incapable, alone, paralyzed, fatigued, useless, inferiors, vulnerable, empty, forced, hesitant, despair, frustrated, distressed, woeful, pathetic, tragic, in a stew, dominated

Indifferent	Insensitive, dull, nonchalant, neutral, reserved, weary, bored, preoccupied, cold, disinterested, lifeless
Afraid	Fearful, terrified, suspicious, anxious, alarmed, panic, nervous, scared, worried, frightened, timid, shaky, restless, doubtful, threatened, cowardly, quaking, menaced, wary
Hurt	Crushed, tormented, deprived, pained, tortured, dejected, rejected, injured, offended, afflicted, aching, victimized, heartbroken, agonized, appalled, humiliated, wronged, alienated
Sad	Tearful, sorrowful, pained, grief, anguish, desolate, desperate, pessimistic, unhappy, lonely, grieved, mournful, dismayed

Table 2: The 16 psycho-linguistic features subset (PsychPage, 2014).

A small number of 4 syntactic features (Lines 26 to 29 in Table 1) was also chosen to reflect the writing style of the author of the text at the sentence level and they are expected to play an important role in the gender identification of the author of the text, since women and men appear to make use of punctuation in different ways (Argamon *et al.,* 2003; and Sterkel, 1988). Also, 9 gender-preferential language features were included to measure the frequent use of emotionally intensive adverbs and affective adjectives such as really, lovely, adorable, marvellous, aggressive, etc., that could help with a prediction of the gender of a person, should it be accepted, bearing in mind the *caveat* [1] already established earlier in the paper, that the tendency to use emotionally intensive words is a more frequent feature of women's writing in social networks, and that the tendency to use singular pronouns and directive sentences is a more frequent feature of men's writing in social networks (Argamon *et al.,* 2009; and Cheng *et al.,* 2011). As it can be seen in Lines 30 to 38 of Table 1, the frequency of these function words is measured through the presence of the included suffixes, e.g. -able, -ive, -less, etc.

Finally 385 function words were included (see Appendix) for the crucial role they are perceived to play in distinguishing writing styles across the gender divide (Chung and Pennebaker, 2007; and Newman *et al.*, 2008). As Lines 39 to 423 of Table 1 show, these words were clustered into 6 groups (articles; pronouns; auxiliary verbs; conjunctions; interjections; and adverbs and prepositions) and the frequency of each function word in a user's text is calculated by dividing the number of appearances of a word in the text by the total number (N) of words in the text sample.

2.3 Part-of-speech tagging

Before extracting function words from the text produced by the social networks users, a *part-of-speech tagging* (Chamiak *et al.*, 1993) is performed by training a part-of-speech tagger. Part-of-speech tagging is a process whereby tokens are sequentially labelled with syntactic labels, such as nouns, prepositions, adjectives, possessive personal pronouns, possessive adverbs, coordinating conjunctions, verbs, etc. A part-of-speech tagger assigns a tag (i.e., a short coded description of the part-of-speech) to every token in the text. To be more precise, a part-of-speech tagger assigns the most likely tag to every token in the text, using a tagger model (the eight standard part-of-speech tags are: adjectives, adverbs, conjunctions, determiners, nouns, prepositions, pronouns and verbs). The purpose of a tag is to provide the grammatical class of each word in the text and some predictive features indicating the behaviour of other words in the textual environment. Unfortunately, since some words in a text may come under more than one syntactic label, a simple check in a dictionary is not an option for the part-of-speech tagger. Therefore the most likely part-of-speech for every word has to be chosen (Chamiak, 1996). There are many different ways of performing part-of-speech tagging, and the best tag set depends on the application. Therefore, the tag set is predefined by an appropriately selected training data set which in the proposed model is the Brown Corpus proposed by Francis and Kucera, 1964. The Brown Corpus is a well-known old-school corpus of English textual material, one million words in length, made up of 500

samples of 2,000 or more words each, from randomly picked works published in the United States in 1961. This paper has chosen on the basis of this dataset a Hidden Markov model (Kupiec, 1992) to be used for performing the part-of-speech tagging. The Hidden Markov model-based tagger will use stochastic methods and probabilities to tag the words of a sentence. Therefore, the success of this scheme will be significantly affected by the selection of an appropriate training dataset to accurately construct the probability estimations for the proposed model. This selection process is described below.

3. Architectural and Implementation Issues

This section will start with a survey of the selection process of the training data that is required to strengthen the robustness of the proposed gender classification model. The section will continue by presenting the implementation of a prototype gender inference system which is based on the proposed architecture and by means of which the efficiency of the new methodology will be put to the test and verified.

3.1 Training Datasets

The Twitter Training Dataset: To obtain a training dataset of tweets, the Twitter platform was crawled randomly and 10,000 tweets written in English were collected, i.e, 5,000 tweets written by women and 5,000 written by men. The gender of the authors of the tweets was subsequently verified manually by means of external information gathered on the web (photos, personal home pages, verified presence in other online social network platforms, etc.). The dataset was used by the text mining module to train the SVM classifier that predicts the gender of Twitter's users on the basis of the textual content of their tweets. The tweets in the training dataset were wiped clean of URLs as well as of topics (“#topic”) and of Twitter usernames (“@name”) in order to obtain a suitable corpus of text samples for the classification process.

The Reuters-21578 Newsgroup Dataset: The Reuters-21578 dataset, proposed by Carnegie Group Inc. and Reuters Ltd., 1987, consists of thousands of articles which cover a range of international newswire and vary from hundreds to thousands of words in length. The articles were sorted into categories according to the gender of the journalists and then 1,000 articles were randomly picked: 500 articles by female and 500 articles by male authors. The dataset was selected to train a SVM classifier that would predict the gender of LinkedIn users from the textual content of their profile's attribute summary. In order not to put at risk the integrity of the accuracy of the proposed gender classifier, the decision was taken to exclude from the dataset the articles that contained several quotes from texts with different authors.

One might well expect that the most successful strategy for the training of the LinkedIn SVM classifier might be the crawling of the LinkedIn platform to collect some thousands of LinkedIn user summaries, the gender of the authors of which would have been verified by human inspection. However, opting for another solution instead, this paper suggests that the Reuters-21578 dataset should be selected which, in terms of preparation and implementation, offers a much faster and simpler solution than any LinkedIn sample dataset that would obviously require excessively time-consuming manual gender verification. The experimental study indicates a perfect correspondence between the neutral character of the descriptive linguistic content that appears in the Reuters articles and the linguistic content found in the LinkedIn summary by professionals describing themselves.

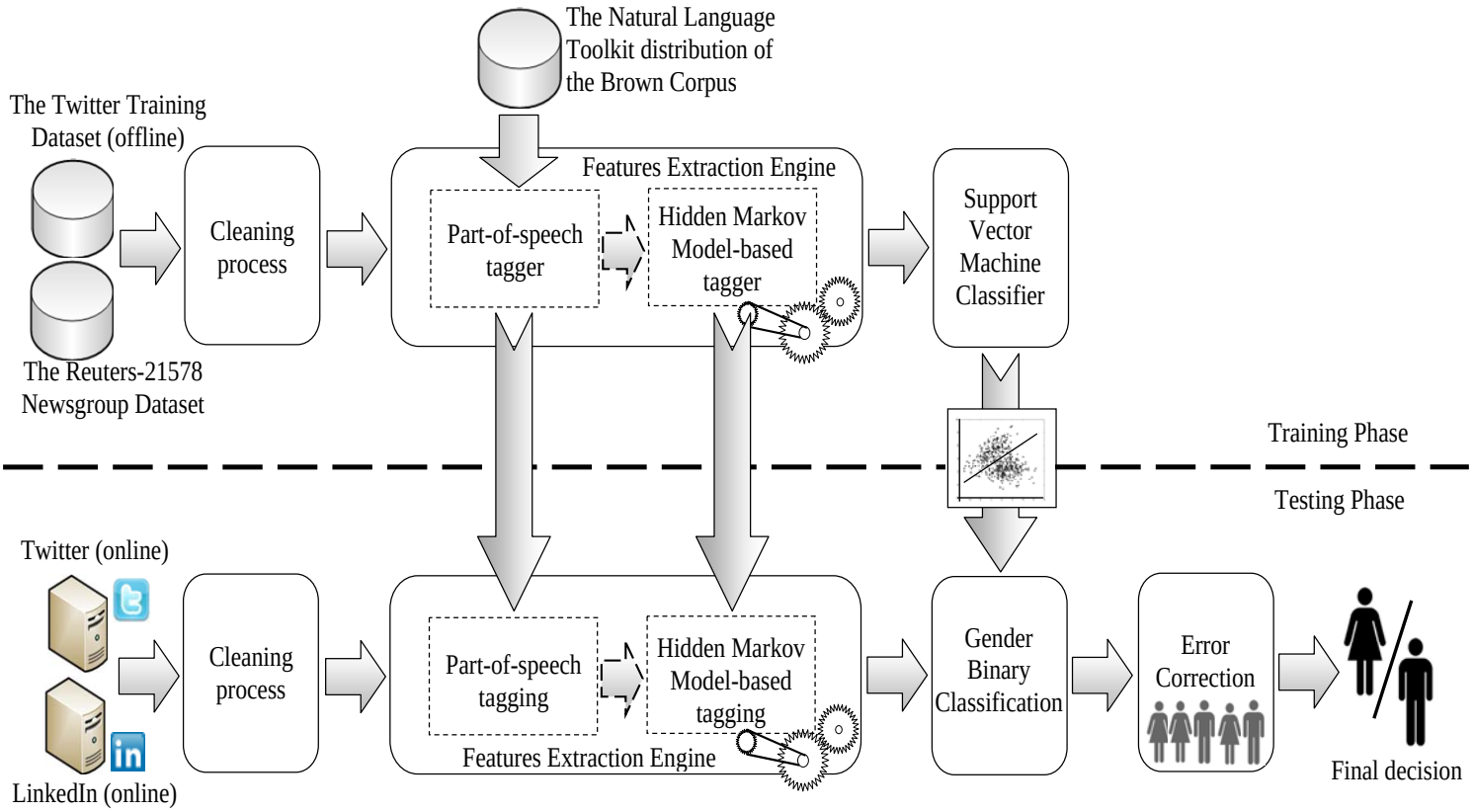

Figure 1: The proposed system's architecture.

3.2 The proposed architecture

As Figure 1 illustrates, the system's Data Cleaning module is responsible for cleaning the input (training and testing) data by removing unnecessary items that do not contribute to the part of speech tagging and to the classification tasks, like for example, usernames, hash tags, URLs, etc.

For each text sample, the Features Extraction Engine computes the 423-dimensional numerical vector that has been specified in Section 2. This module incorporates two processing tools: a part-of-speech tagger and a Hidden Markov model-based tagger. The part-of-speech tagger is constructed on the basis of the LingPipe Toolkit [LingPine, 2014] which provides a part-of-speech tag parser for the open-source Natural Language Toolkit distribution (NLTK, Bird, 2006) of the Brown Corpus. This training text is formatted in pure ASCII, with sentences delimited, tokens delimited and tags separated from

tokens by a forward slash. This corpus is used to train our part-of-speech tagger and then to load it into the probabilistic (stochastic) Hidden Markov model-based tagger in order to extract some of the features of interest (only for the features related to function words), which will form, together with some style markers, the 423-feature set that was presented in Section 2.

Regarding the training samples, the produced feature vectors are then used to build an efficient classification module which is comprised of a linear SVM classifier for Twitter and an analogues classifier for LinkedIn, both constructed using the LingPipe Toolkit. These two binary classifiers will then comprise the 'Gender Binary Classification' component of the system that will efficiently determine the category <female> or <male> of any relevant input text data and would thus predict the gender of its author.

On LinkedIn one single testing text sample is available for each targeted author (i.e. the summary field) while on Twitter more than one tweet for testing might exist, thus, the proposed algorithm will generate a prediction for each sample and then the algorithm will deploy an error correcting mechanism to select the most likely gender of that author. For example if ten tweets are available, and seven of these point to the author's as female while the three other tweets point to the author as male, the system can use a majority voting scheme to produce the conclusion that the author is most likely female.

4. Evaluation

The selected features-set of the designed methodology together with the quality of the training data are the two most critical factors for building a classification tool that can predict effectively the two classes <female> and <male> of the binary gender attribute in a collection of written texts. The results obtained by the experimental evaluation of the implemented prototype gender classification tool will be reported below.

4.1 Experimentation Setup

To evaluate the accuracy of the gender prediction tool, public English data samples from LinkedIn and Twitter platforms were used. More specifically, the summary attribute of 1,000 randomly selected users' profiles on LinkedIn and 1,000 English tweets of randomly selected users' profiles on Twitter were crawled and obtained, excluding tweets containing only hyperlinks and tweets containing one single word of text. The gender of the authors of the test data was also manually examined and validated [2]. None of the tweets of these authors appear into the Twitter Training Dataset.

In the testing phase in Twitter, *Accuracy* was defined as:

$$Accuracy = \frac{\text{number of tweets the author's gender of which was correctly inferenced}}{\text{total number of tweets in the dataset}}$$

and in the testing phase in LinkedIn, *Accuracy* was defined as:

$$Accuracy = \frac{\text{number of LinkedIn summaries the author's gender of which was correctly inferenced}}{\text{total number of LinkedIn summaries in the dataset}}$$

where the gender of an author is correctly predicted if it is validated according to the corresponding manual human inspection.

4.2 Performance Evaluation

On the basis of the 1,000 testing tweets from Twitter, as illustrated in Figure 2 922 correct gender predictions were obtained, which meant a 92.2% accuracy rate, and a 7.8% error rate, for the proposed Twitter SVM classifier.

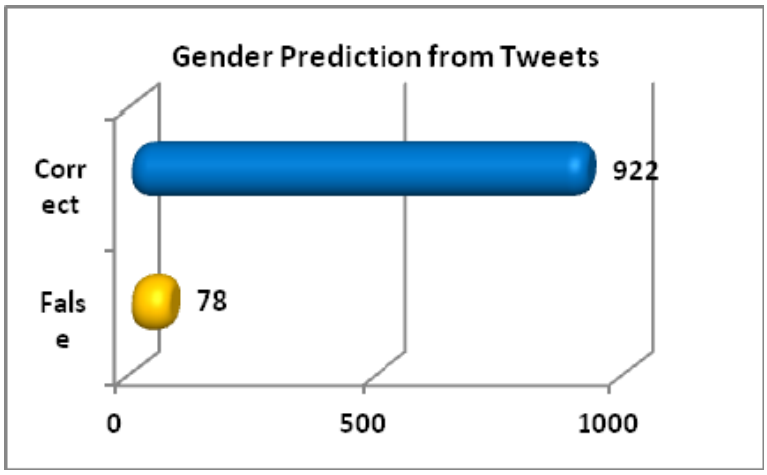

Figure 2: Gender prediction based on Tweets.

Therefore in 78 cases only the gender prediction carried out by the classifier was not successful. These cases were manually examined and it was discovered that the concerned tweets were of poor content, which means that they did not provide sufficient useful written textual material to deal with. The significance of these results resides in the fact that Twitter is a micro-blogging service that offers small messaging possibilities to a maximum length of 140-characters, and often very poor in respect of content.

On LinkedIn the gender classifier could make use of a smaller features-set of lesser dimensions since it was not necessary to include in the set the features of Table 1 that are related to emotional language because both the training and the testing datasets consisted of texts in descriptive linguistic

content of a neutral character. On the basis of this modification, as Figure 3 illustrates, the proposed model predicted correctly the gender of 984 LinkedIn users which means a 98.4% accuracy, and a 1.6% error rate for the corresponding SVM classifier.

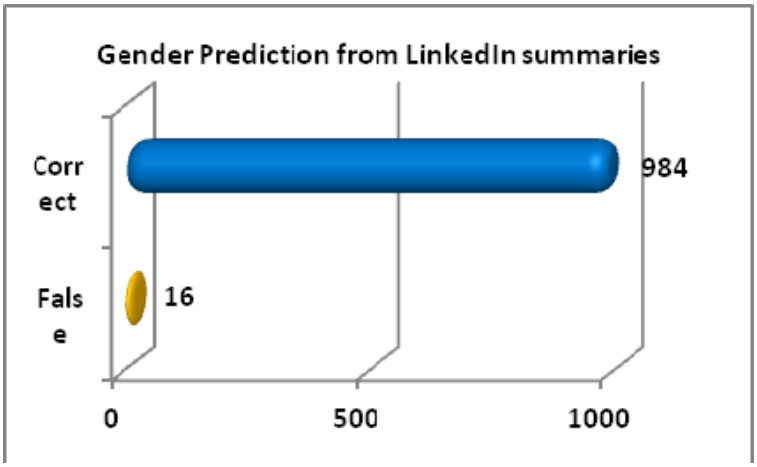

Figure 3: Gender prediction based on LinkedIn Summaries.

4.3 Discussion

The promising experimental results in addressing the problem of user gender prediction from tweets in Twitter and LinkedIn summaries on the basis of the proposed model indicate that the prediction of the gender of the users on social networks that do not, as a rule, reveal such items of personal information, is indeed possible to achieve. The proposed research indicates that, by selecting a sophisticated set of features and a suitable training dataset, it is possible to construct a robust SVM gender classification model to accomplish this task.

It is also important to highlight that, in the case of LinkedIn, the Reuters-21578 training dataset consists of news articles, while the testing dataset consists of user profiles summaries which are different from the perspective of notional structure and focus, as well as being of a length that is much shorter

than the average length of the news articles. It might therefore be argued that the Reuters-21578 dataset is far from a perfectly fitting training dataset for the targeted data collection. However, the choice made by the authors appears to be vindicated since the experimental results proved very promising and even more successful than the corresponding results with Twitter.

On the other hand, the experimental results on Twitter clearly indicate that the task of achieving gender inference using a single tweet is more difficult than when a LinkedIn summary is used. The explanation for this has to do with the average number of words contained in the testing sample in every individual case. It is self-evident that the rate of accuracy increases in relation to the number of words in the written text sample: the more words there are, the more valuable information will be contained therein for the SVM classifier. Since the average length of a LinkedIn's summary is much greater than the average length of a tweet, it follows that the success rate in user's gender inference in LinkedIn will consistently be higher.

5. Related Work

Over the last few years and while online social networking platforms were making tremendous gains in popularity with both common and malicious users, and as a result of the availability of the massive amounts of sensitive data about their users which, by virtue of their very nature, they have been making publicly available, the emphasis of research has been directed to issues of information leakage and disclosure on social networks (Chew *et al.,* 2008; Carter and Mistree, 2009; Mislove *et al.,* 2010; Xu *et al.,* 2008; and Zhelev and Getoor, 2009).

The privacy of online social networks is facing a threat of monumental importance: the attribute disclosure threat. Attribute disclosure occurs when the attacker is able to determine the value of a user's attribute that the user wishes to keep undisclosed. The works carried out in Lanza and Svendesen, 2007 and in Zhelev and Getoor, 2009 focus on using public profiles, friendship links and group memberships to

infer a user's undisclosed private information. The malicious party may efficiently predict the age group or the mother tongue or the political affiliation of a user through an examination of similar attributes amongst the user's friends. The complementary work carried out in Mislove *et al.*, 2010 explores how a malicious party can manage to predict a private attribute of a targeted individual by means of combining the social network graph with the similar attribute of a number of other users who are not directly connected to the targeted user. In their experimental study the authors are able, with as little as 20% of users with known attributes, to infer the remaining users' common attributes at an 80% rate of accuracy. The practical conclusion to which these three works of research lead us is that a user's privacy can be easily compromised by exploiting knowledge obtained through her/his online social relationships. Similarly, Xu *et al.*, 2008 proposes an algorithm to infer a user's private attribute by mining only her/his social relations. Also, He *et al.*, 2006 utilizes Bayesian networks to measure the probability that a specific user may have a given attribute value. As demonstrated in the above article, by using the LiveJournal online social network, inferring with a high rate of accuracy a user's attribute is achievable even in online communities where the majority of users keep their personal data undisclosed.

The research on attribute disclosure studied thoroughly the inference of various types of users' personal attributes, such as age, gender, profession, religious persuasion, etc. The remainder of the article will concentrate mainly on the issue of gender inference since gender is the attribute which represents the focus of this work. In Cheng *et al.*, 2011 the authors conducted a series of experiments with three different statistical and probabilistic machine learning classification techniques (SVM, Bayesian Logistic Regression and Decision Tree) to yield a maximum gender prediction accuracy of 85.1%. In Pennebaker *et al.*, 2011 the authors propose a text categorization approach for the prediction of age group and gender on a corpus of chat texts which was compiled from the Belgian social networking site NetLog. The selected features-set is limited to token and character-based features: words unigrams, bigrams, and trigrams, and also character bigrams, trigrams and tetra grams. The paper deploys a SVM classification model that yields an accuracy score of 88.8%. In Filippova, 2012 the author focuses on the

gender prediction of YouTube users, on the basis of comments and the affiliation graph of users and videos, reporting accuracy levels hovering around the 90% mark, but depending on the users' age group.

In Rao *et al.*, 2010 and Rao and Yarowsky, 2010 the authors experiment with stacked SVMs-based classification algorithms over a rich set of features, both lexical (n-gram-based) and sociolinguistic features. The gender prediction accuracy score that was achieved was 72.33%, based on a training dataset of 500 Twitter accounts. Similarly, in Burger *et al.*, 2011 the authors provide experiments with a variety of different machine learning algorithms, including SVMs, Naïve Bayes and Balanced Winnow to build gender classification models. The deployed classifier relying only on a single tweet performed at a 76% accuracy level. According to the paper, human prediction performance was not too far off this mark, in comparison, scoring a 68.7% average.

The algorithms presented in Miller *et al.*, 2012 rely on the perception and naïve Bayes stream machine learning algorithms and make use of a n-grams features-set to represent tweets training and testing data. The perception algorithm achieves the highest accuracy level among all, with a maximum scoring of 99.3% when the tweets length is of at least 75 words. In Deitrick *et al.*, 2012 the neural network machine learning algorithm Modified Balanced Winnow, with an extensive list of 53 n-gram features and only a limited dataset of 1,484 training tweets, achieved a 98.5% level of accuracy. A similar experiment is carried out in Fink *et al.*, 2012 to infer the gender of Twitter users by implementing a SVM classification model based on word unigrams, hash tags and psychometric properties derived from a text analysis application called Linguistic Inquiry and Word Count or LIWC (Pennebaker *et al.*, 2007). The reported accuracy is 80% in predicting the users' gender by using the unigrams features alone.

Bamman *et al.*, 2012 and Zamal *et al.*, 2012 study the influence of a Twitter user's immediate neighbours' attributes in predicting a similar private attribute of the user. In the case of gender the accuracy of their prediction model is of 88% and 80.2% respectively. Bergsma *et al.*, 2013 examines clusters of Twitter users based on preferences and location and then uses this information with the

gender prediction task; the accuracy of the proposed method is of 90.2%. Also, with the manual inspection of 120 Twitter users' profiles carried out by human experts in the domain of language analysis, the article reports an accuracy rate in gender prediction of, at the most, 88.3%. Liu and Ruths, 2013 incorporates the user's self-reported name into the Twitter gender classifier and achieves a prediction accuracy of 87.1%. Finally, Liu *et al.*, 2012 identifies three Twitter accounts dedicated to broadcasting information about traffic, public transportation issues, and cycling in Toronto and then uses the community of followers of these accounts as a test sample of users for demographic inference. The proposed specialized model for these types of commuter populations yields a maximum gender prediction accuracy rate of 86.8%.

This article focuses on the user's gender prediction problem in online social networks. The proposed classification model is implemented and applied to Twitter and LinkedIn (the platforms of which do not reveal information about the gender attribute) yielding a prediction accuracy rate of up to 98.4%. The research effort that stands closest to the proposed work is that of Argamon *et al.*, 2009, which, however, relies on a Bayesian Multinomial Regression probabilistic machine learning algorithm and uses blog posts as training data to learn the proposed classification model for predicting the profile of an anonymous author. The gender classification module in this earlier article achieves 76.1% in accuracy.

Table 3 summarizes the performance achievements of previous work in terms of gender inference accuracy (%) for Twitter users. For every case, the table shows the best reported performance accuracy levels, even if this is achievable only under significantly restricted conditions (such as, for example, the accuracy achievements of Miller *et al.*, 2012, which scores up to 99.3% but only when the length of the tweets under examination is of at least 75 words). For a comparison with the proposed work, the accuracy of the model proposed in this paper appears in the bottom line of the table. The proposed scheme outperforms most of the earlier work in predicting the users' gender in Twitter and, to the best of the authors' knowledge, this scheme is the first to address the gender discrimination problem in LinkedIn (the abbreviation 'N/A' in Table 3 stands for the phrase 'Not Addressed').

Gender prediction model	Accuracy in Twitter	Accuracy in LinkedIn
Bamman *et al.*, 2012	88.0%	N/A
Bergsma *et al.*, 2013	90.2%	N/A
Burger *et al.*, 2011	76.0%	N/A
Deitrick *et al.*, 2012	98.5%	N/A
Fink *et al.*, 2012	80.6%	N/A
Human inspection Burger *et al.*, 2011	68.7%	N/A
Human expert inspection Bergsma *et al.*, 2013	88.3%	N/A
Liu *et al.*, 2012	86.8%	N/A
Liu and Ruths, 2013	87.1%	N/A
Miller *et al.*, 2012	99.3%	N/A
Rao *et al.*, 2010; and Rao and Yarowsky, 2010	72.3%	N/A
Zamal *et al.*, 2012	80.2%	N/A
This paper	**92.2%**	**98.4%**

Table 3: Accuracy provided by gender inference models in Twitter and LinkedIn.

6. Conclusion

Social networking services are the spearhead of the new digital economy in the dot com world, offering a new field for applications and innovative ideas for added-value services that generate new markets and produce more profits for the capitalist investors. Intelligent and perceptive analysis of online social network data is a valuable resource for the whole spectrum of different parties from advertisers to intelligence services, as it can lead to new insights in both commercial and governance settings, and to competitive advantages. In this study a new approach for inferring the users' gender in online social networks is proposed. A new text mining module that combines a Hidden Markov Model part-of-speech tagger with a SVM linear binary classifier is constructed to perform gender predictions from online social networks textual data. The experimentation with Twitter has shown that the proposed scheme can successfully determine the users' gender attribute with higher levels of accuracy than most of the other related efforts in the field until now. The corresponding study on LinkedIn using profile summaries written in English produced a remarkably robust accuracy performance of 98.4%, without any restriction having being imposed on the tested data. It should also be noted that, to the best of the authors' knowledge, no research work has considered, to-date, the issue of gender prediction in the LinkedIn social network setting and that an advantage of the proposed gender inference model is that it can easily be fine-tuned to function efficiently in any online social network and in any textual language since most of its selected measures are not based on n-gram features.

There still is a wide expanse of unexplored ground to be covered by further research and development in the domain of predicting undisclosed attributes or of inferring information about users in online social networks. The authors’ intention is to extend the proposed model to the inference of other hidden attributes and to develop it to take into consideration new factors and features across textual data. It is also planned to integrate many separate modules for attributes disclosure in order to construct an extended hybrid authorship profiling engine that could establish links with any information that may be of

value for the purpose of targeting individuals from any discovered online footprint and any trustworthy source.

About The Authors

Mr. Athanasios Kokkos received a B.Sc. in Information Technology from the Technological Educational Institute of Thessaloniki and a M.Sc. in Information and Communication Systems Security from the University of the Aegean, in Greece. Currently he is a PhD Candidate in the University of the Aegean. His research interests include data security and social networks.

E–mail: ath.kokkos [at] aegean [dot] gr

Dr. Theodoros Tzouramanis received a 5-year B. Eng. (1996) in Electrical and Computer Engineering and a PhD (2002) in Informatics, both from the Aristotle University of Thessaloniki, in Greece. Currently he is an Assistant Professor and the Head of the Database Laboratory in the Department of Information and Communication Systems Engineering of the University of the Aegean. His research interests include access methods and query processing algorithms for temporal, spatial, image, multimedia databases; databases for time-evolving spatial data; databases for GIS; and, data privacy, security and forensics.

Web: http://www.icsd.aegean.gr/ttzouram/

Direct comments to: ttzouram [at] aegean [dot] gr

Notes

[1] *Caveat:* It is important to note that this study does not take a stand with regard to the debate around gender issues and the authors are fully aware that they could be seen to tread on shifty ground. The

notion of gender as a women-men binary attribute in the stereotypically traditional way provides a suitable platform to demonstrate the approach taken by the work carried out in this context of research. The terms woman/female and man/male are used interchangeably in the remainder of the paper.

[2] It should be noted that the authors' manual inspection of the data corresponded to various gender cues to determine what they felt was the correct gender of the users in relation to the binary definition of this attribute.

References

S.Argamon, M.Koppel, J.Fine and A.R.Shimoni, 2003. "Gender, genre, and writing style in formal written texts", *Text - Interdisciplinary Journal for the Study of Discourse*. Vol.23, No.3, pp.321-346.

S.Argamon, M.Koppel, J.W.Pennebaker, and J.Schler, 2009. "Automatically profiling the author of an anonymous text", *Communications of the ACM*, Vol.52, No.2, pp.119-123.

J.D.Burger, J.Henderson, G.Kim, and G.Zarrella, 2011. "Discriminating Gender on Twitter", In Proceedings of the *Conference on Empirical Methods in Natural Language Processing*, pp.1301-1309.

S.Bergsma, M.Dredze, B.Van Durme, T.Wilson, and D.Yarowsky, 2013. "Broadly improving user classification via communication-based name and location clustering on twitter". In Proceedings of the *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), pp.1010-1019.

I.K.Broverman, S.R.Vogel, D.M.Broverman, F.E.Clarkson, and P.S.Rosenkrantz, 1972. "Sex-Role Stereotypes: A Current Appraisal1", *Journal of Social issues*, Vol.28, No.2, pp.59-78.

D.Bamman, J.Eisenstein, and T.Schnoebelen, 2012. "Gender in twitter: Styles, stances, and social networks". arXiv preprint arXiv:1210.4567.

S.Bird, 2006. "NLTK: the natural language toolkit". In Proceedings of the *COLING/ACL on Interactive presentation sessions* (COLING-ACL '06), pp.69-72.

E.Charniak, C.Hendrickson, N.Jacobson, and M.Perkowitz, 1993. "Equations for part-of-speech tagging", In Proceedings of the *AAAI*, pp.784-789.

M.Corney, O.de Vel, A.Anderson, and G.Mohay, 2002. "Gender-Preferential Text Mining of E-mail Discourse", In Proceedings of the *18th Annual Computer Security Applications Conference*, pp. 282-289.

N.Cheng, R.Chandramouli, and K.P.Subbalakshmi, 2011. "Author Gender Identification from Text", *Digital Investigation*, Vol.8, No.1, pp.78-88.

M.Chew, D.Balfanz, and B.Laurie, 2008. "(Under) mining Privacy in Social Networks", In Proceedings of the *Web 2.0 Security and Privacy Workshop* (W2SP).

E.Charniak, 1996. *Statistical language learning*, MIT Press.

C.Chaski, 2005. "Who's at the keyboard? Authorship attribution in digital evidence investigations", *International Journal of Digital Evidence*, Vol.4, No.1 pp.1-13.

J.Carter, and B.F.T.Mistree, 2009. "Gaydar: Facebook Relationships expose sexual orientation", *First Monday*, Vol 14, No.10.

C.K.Chung, and J.W.Pennebaker, 2007. "The psychological functions of function words". In K. Fiedler (Ed.), *Social Communication*, pp.343-359. Psychology Press.

Carnegie Group, Inc. and Reuters Ltd, 1987. "Reuters-21578". Address to download: http://www.daviddlewis.com/resources/testcollections/reuters21578/, accessed on January 2014.

M.Crawford, 1995. *Talking difference: On gender and language*, Sage.

J.Diederich, J.Kindermann, E.Leopold, and G.Paass, 2000. "Authorship attribution with support vector machines", *Applied Intelligence*, Vol.19, pp.109-123.

W.Deitrick, Z.Miller, B.Valyou, B.Dickinson, T.Munson, and W.Hu, 2012. "Gender Identification on Twitter Using the Modified Balanced Winnow", *Communications and Networks*, Vol.4, No.3, pp.189-195.

A.H.Eagly, 1987. *Sex differences in social behavior: A social-role interpretation*, Psychology Press.

A.H.Eagly and V.Steffen, 1984. "Gender stereotypes stem from the distribution of women and men into social roles", *Journal of Personality and Social Psychology*, Vol.46, No.4, pp.735-754.

K.Filippova, 2012. "User Demographics and Language in an Implicit Social Network", In Proceedings of the *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.1478-1488.

W.N.Francis and H.Kucera, 1964. "Brown Corpus Manual", *Brown University*. Address to download: http://www.hit.uib.no/icame/brown/bcm.html. Address to download the Brown Corpus: http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml, accessed on January 2014.

C.Fink, J.Kopecky, and M.Morawskib, 2012. "Inferring Gender from the Content of Tweets": A Region Specific Example, In Proceedings of the *Sixth International AAAI Conference on Weblogs and Social Media*, pp.459-462.

A.G.Greenwald and M.R.Banaji, 1995. "Implicit social cognition: attitudes, self-esteem, and stereotypes", *Psychological Review*, Vol.102, No.1, pp.4-27.

J.He, W.W.Chu, and Z.Liu, 2006. "Inferring Privacy Information from Social Networks", In Proceedings of the *Intelligence and Security Informatics*, pp.154-165.

T. Joachims, 1998. "Text categorization with Support Vector Machines: Learning with many relevant features", *LNCS*, Vol.1398, Springer Berlin Heidelberg, pp.137-142.

J.Kupiec, 1992. "Robust part-of-speech tagging using a hidden Markov model", *Computer Speech & Language*, Vol.6, No.3, pp.225-242.

LingPipe Homepage, 2014. "BrownPosParser", Address to download: http://alias-i.com/lingpipe/demos/tutorial/posTags/read-me.html, accessed on January 2014.

LinkedIn Developers, 2014. "LinkedIn Profile Fields", Available at: http://developer.linkedin.com/documents/profile-fields, accessed on January 2014.

W.Liu, and D.Ruths, 2013. "What's in a Name? Using First Names as Features for Gender Inference in Twitter". In 2013 *AAAI Spring Symposium Series*, pp.10-16.

E.Lanza, and B.A.Svendsen, 2007. "Tell me who your friends are and I might be able to tell you what language (s) you speak: Social network analysis, multilingualism, and identity", *International Journal of Bilingualism*, Vol.11, No.3, 275-300.

W.Liu, F.Al Zamal, and D.Ruths, 2012. "Using social media to infer gender composition of commuter populations". In Proceedings of the *When the City Meets the Citizen Workshop* (*International Conference on Weblogs and Social Media*).

A.Mislove, B.Viswanath, K.P.Gummadi, and P.Druschel, 2010. "You Are Who You Know: Inferring User Profiles in Online Social Networks", In Proceedings of the *Third ACM international conference on Web search and data mining*, pp.251-260.

Z.Miller, B.Dickinson, and W.Hu, 2012. "Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features", *International Journal of Intelligence Science*, Vol. 2, pp.143-148.

I.B.Mauss and M.D.Robinson, 2009. "Measures of emotion: A review", *Cognition and Emotion*, Vol.23, No.2, pp.209-237.

M.L.Newman, C.J.Groom, L.D.Handelman, and J.W.Pennebaker, 2008. "Gender differences in language use: An analysis of 14,000 text samples", *Discourse Processes*, Vol.45, No.3, pp.211-236.

J.W.Pennebaker, C.K.Chung, M.Ireland, A.Gonzales, and R.J.Booth, 2007. "The development and psychometric properties of LIWC2007 (Software manual)", Austin, TX, LIWC. Net.

C.Peersman, W.Daelemans, and L.Van Vaerenbergh, 2011. "Predicting Age and Gender in Online Social Networks", In Proceedings of the *3rd international workshop on Search and mining user-generated contents*, pp.37-44.

PsychPage, 2014. "List of Feeling Words", Address to download: http://www.psychpage.com/learning/library/assess/feelings.html, accessed on January 2014.

D.Rao, and D.Yarowsky, A.Shreevats, and M.Gupta, 2010. "Classifying Latent User Attributes in Twitter", In Proceedings of the *2nd international workshop on Search and mining user-generated contents*, pp.37-44.

D.Rubin and K.Greene, 1992. "Gender-typical style in written language", *Research in the Teaching of English*, Vol.26, No.1, pp.7-40.

D.Rao, and D.Yarowsky, 2010. "Detecting Latent User Properties in Social Media", In Proceedings of the *NIPS MLSN Workshop*.

S.A.Shields, D.N.Garner, B.Di Leone, and A.M.Hadley, 2006. "Gender and Emotion", In *Handbook of the sociology of emotions*, Springer US, pp. 63-83.

K.S.Sterkel, 1988. "The Relationship Between Gender and Writing Style in Business Communications", *Journal of Business Communication*, Vol.25, No.4, pp.17-38.

W.Xu, X.Zhou, and L.Li, 2008. "Inferring Privacy Information via Social Relations", In Proceedings of the *IEEE 24th International Conference on Data Engineering Workshop* (ICDEW'08), pp.525-530.

E.Zhelev, and L.Getoor, 2009. "To Join or Not To Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles", In Proceedings of the *18th International Conference on World Wide Web* (WWW).

F.Al Zamal, W.Liu, and D.Ruths, 2012. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors". In Proceedings of the *Sixth International AAAI Conference on Weblogs and Social Media*, pp.387-390.

Appendix: Function Words

Article words

a, an, the

Pronoun words

all, everybody, his, most, other, that, what, your, another, everyone, I, much, others, theirs, whatever, yours, any, everything, it, myself, ours, them, which, yourself, anybody, few, its, neither, ourselves, themselves, whichever, yourselves, anyone, he, itself, no, one, several, these, who, anything, her, little, nobody, she, they, whoever, both, hers, many, none, some, this, whom, each, herself, me, nothing, somebody, those, whomever, each, other, him, mine, one, someone, us, whose, either, himself, more, one another, something, we, you.

Auxiliary-verbs

are, can, didn't, hadn't, haven't, might, shouldn't, won't, aren't, cannot, do, 'd, 've, mightn't, was, 'll, ain't, can't, don't, has, is, mustn't, wasn't, would, 're, could, does, hasn't, isn't, shall, were, wouldn't, be, couldn't, doesn't, 's, 's, shan't, weren't, 'd, been, did, had, have, may, should, will.

Conjunction words

and, or, though, now, that, if, while, in order that, in case, because, yet, unless, even though, now that, whereas, even if, nor, so, when, although, only if, whether or not, until.

Interjection words

adios, bah, dear, ha-ha, howdy, oops, tush, whoosh, ah, begorra, doh, hail, hoy, ouch, tut, wow, aha, behold, duh, hallelujah, huh, phew, tut-tut, yay, ahem, bejesus, eh, heigh-ho, humph, phooey, ugh, yikes, ahoy, bingo, encore, hello, hurray, pip-pip, uh-huh, yippee, alack, bleep, eureka, hem, hush, pooh uh-oh, yo, alas, boo, fie, hey, indeed, pshaw, uh-uh, yoicks, all, hail, bravo, gee, hey presto, jeepers creepers, rats, viva, yoo-hoo, alleluia, bye, gee, whiz, hi, jeez, righto, voila, yuk, aloha, cheerio, gesundheit, hip, lo and behold, scat, wahoo, yummy, amen, cheers, goodness, hmm, man, shoo, well, zap, attaboy, ciao, gosh, ho, my, word, shoot, whoa, aw, crikey, great, ho, hum, now, so long, whoopee, ay, cripes, hah, hot dog, ooh, touch, whoops.

Adverbial and prepositional words

aboard, astride, down, of, through, worth, on to, in front of, about, at, during, off, throughout, according to, onto, in lieu of, above, athwart, except, on, till, ahead to, out from, in place of, absent, atop, failing, onto, to, as to, out of, in spite of, across, barring, following, opposite, toward, aside from, outside of, on account of, after, before, for, out, towards, because of, owing to, on behalf of, against, behind, from, outside, under, close to, prior to, on top of, along, below, in, over, underneath, due to, pursuant to, versus, alongside, beneath, inside, past, unlike, except for, regardless of, concerning, amid, beside, into, per, until, far from, subsequent to, considering, amidst, besides, like, plus, up, in to, as far as, regarding, among, between, mid, regarding, upon, into, as well as, apart from, amongst, beyond, minus, round, via, inside of, by means of, around, but, near, save, with, instead of, in accordance with, as, by, next, since, within, near to, in addition to, aslant, despite, notwithstanding, than, without, next to, in case of.