Inf Syst E-Bus Manage (2017) 15:117–138
DOI 10.1007/s10257-016-0312-0

ORIGINAL ARTICLE

Gender classification of microblog text based on authorial style

Shubhadeep Mukherjee[1] · **Pradip Kumar Bala**[1]

Received: 14 March 2015 / Revised: 12 January 2016 / Accepted: 15 February 2016 /
Published online: 2 March 2016

Abstract Gender profiling of unstructured text data has several applications in areas such as marketing, advertising, legal investigation, and recommender systems. The automatic detection of gender in microblogs, like twitter, is a difficult task. It requires a system that can use knowledge to interpret the linguistic styles being used by the genders. In this paper, we try to provide this knowledge for such a system by considering different sets of features, which are relatively independent of the text, such as function words and part of speech n-grams. We test a range of different feature sets using two different classifiers; namely Naïve Bayes and maximum entropy algorithms. Our results show that the gender detection task benefits from the inclusion of features that capture the authorial style of the microblog authors. We achieve an accuracy of approximately 71 %, which outperforms the classification accuracy of commercially available gender detection software like Gender Genie and Gender Guesser.

Keywords Text mining · Twitter · Natural language processing · Gender classification · Knowledge discovery · Supervised learning · Artificial intelligence · Business intelligence

Electronic supplementary material The online version of this article (doi:10.1007/s10257-016-0312-0) contains supplementary material, which is available to authorized users.

✉ Shubhadeep Mukherjee
shubhadeep.mukherjee13fpm@iimranchi.ac.in

Pradip Kumar Bala
pkbala@iimranchi.ac.in

[1] Information System and Analytics Area, Indian Institute of Management Ranchi, Ranchi, Jharkhand 834001, India

1 Introduction

With the rapid growth of social media there has been an unprecedented increase in the amount of user-generated data. The open availability of this data on public networks, particularly on various social networking sites, provides researchers and organizations with ample opportunity to study patterns present in the linguistic styles of the informal language and utilize it to reap the maximum benefit for organizations. Studying the patterns of this nouveau language can provide a great deal of insight for businesses, researchers, organizations combating cybercrime, understanding social opinions, etc.

One such study is the gender classification of unstructured text data. It attempts to learn the subtle variations in the writing styles between the genders by studying the linguistic styles of men and women. Identifying the gender of the author of a given text has been an important classification problem since early 2000. Researchers have studied gender classification of text based on natural language processing (NLP) extensively (Koppel 2002; Argamon et al. 2003; Hota et al. 2006; Mukherjee and Liu 2010; Rao et al. 2010). However, the classification of tweets or microblogs by gender is only now being explored.

The problem of gender classification is of growing importance in the current global climate. Large corporations are interested in knowing what types of people (male or female) like their products based on analysis of blog posts, including microblogs. These reviews are helpful in many commercial domains, such as target advertisement and product development. Likewise, intelligence departments may use gender classification for crime investigation (Peersman et al. 2011). Furthermore, a user's experience with a microblogging service could be significantly improved if information about the demographic attributes or personal interests of particular user, as well as other users of the service, were available. Such information could allow for personalized recommendations of users to follow or user posts to read (Pennacchiotti and Popescu 2011). Additionally, events or topics of interest of the particular gender could be highlighted. Any information that can be gleaned from authorship may have applications across variety of fields; for instance, gender identification has applications in marketing, advertising, legal investigation and understanding social opinions of the genders.

It is essential to differentiate between regular blogging and microblogging. From a philosophical standpoint, blogging sites are not meant for updating friends and acquaintances about one's daily activities. Microblogs, however, are used frequently for such purposes. Microblogs impose restrictions on the number of words or characters one can write per post (e.g. 140 characters for Twitter) but there are no such restrictions on regular blogs. Also, microblogging doesn't necessarily need to have an agenda and can be completely random (e.g. I feel awesome today☺☺, a photograph with a caption etc.), while blogs, on the other hand, are often related to a particular topic of the interest for the author which is usually well described by the author. One can thus say that microblogs can be quite different from regular blogs in style and content. These differences necessitates separate treatment for microblogs.

The classification of microblogs like Twitter is a challenging problem. This is due to the following reasons:

(a) The maximum length of a tweet is 140 characters, which means one would need a much larger training set for the purpose of data training.
(b) There are accidental and purposeful misspellings in tweets, which are almost absent in regular texts (Cooool, goooood etc.).
(c) Part of speech tagging is a daunting task due to the informal nature of the tweets.
(d) Presence of emoticons and acronyms (☺, lol, rofl etc.).

However, the syntactic structure of a tweet in English roughly follows the syntactic structure of an English sentence. An English sentence can be broadly said to consist of two types of words—function words and content words. Function words are words that have little lexical meaning or have ambiguous meaning, and instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Function words are words, such as "the", which have a particular grammatical role but little identifiable meaning (Klammer et al. 2000). On the other hand, content words typically carry semantic content, bearing reference to the world independently of its use within a particular sentence. A word to which an independent meaning can be given, by reference to a world outside a sentence, in which the word may occur (Winkler 2012).

We approach the problem of gender classification of tweets in a way similar to that used in classification of regular text by Argamon et al. (2003, 2007). We extract features based on function words, part of speech n-gram tags and the most popular content words. From these features we expect to get a greater insight into the subtle differences in the types of linguistic features used by the two genders on social media. The most frequent function words and content words, as well as the part of speech n-grams used by either gender in non-formal communication, could be studied to generate new business, research, and social insights. In the extant research literature we have not come across any work where, in the gender classification of microblogs, function words and part of speech n-grams have been used as features.

Our work adds to the existing body of knowledge in the following three ways:

1. As a first in the field, function words and part of speech n-grams have been used as features to classify microblog text. It's a unique effort in the area to establish features that appear innocuous but could be relevant in classification of unstructured data in case of other related classification problems in social media.
2. We improve the classification accuracy by 7 % over the existing gender classification software with the above-mentioned features for a small dataset (3000 tweets). It logically follows that on increasing the dataset size the classification accuracy should further improve. However, we also need to ensure that the improvement achieved should not be at the cost of precision and

recall of the classification job, thus we have incorporated F-measure. F-measure is a well-accepted measurement criteria in data mining research as it does not get affected by class imbalance and provides the harmonic mean for precision and recall.

3. Since, we have focused on using features that are independent of the text, the obtained results are more applicable universally than the results obtained by capturing text based features. A comparative study of the performance of the various features captured in the research has also been done.

2 Related work

Regular text classification for authorship profiling has been addressed as a research problem since early 2000's (Koppel 2002; Argamon et al. 2003). These authors classified formal textbooks based on writing style. They used British National Corpus (BNC) tagged corpora for training features for classification. This was one of the first works in the classification of authorship of formal text based on gender. Subsequent blog classification was attempted by Yan and Yan (2006). The authors used simple word features, background colors, and emoticons to classify text using the Naïve Bayesian algorithm. Zhang and Zhang (2010) captured word features and simple part of speech tags to classify the gender of blog authors. Later, Mukherjee and Liu (2010) used word sequence n-grams and feature selection ensembles for classifying blog text. The authors also compared the classification accuracy of their algorithm against the commercially available software and found better results.

Gender classification of microblog authors is relatively new and is just starting to be explored by researchers. One of the more remarkable works in the field is by Rao et al. (2010), which has captured latent user attributes built on a support vector machine (SVM) based algorithm. The authors used n-gram word features of tweets as gender differentiators for their dataset. They also classified authors based on religious beliefs and political orientation using the same features. Penachiotti and Popescue (2011) used rich linguistic features for classification. They applied a machine learning approach on a comprehensive set of features derived from relevant user information. Alowibdi et al. (2013) used non-textual features like background colors and its combinations to classify twitter user profiles based on gender and got reasonably high accuracy. Miller et al. (2012) used character level n-grams as a feature to classify Twitter text. They applied Naïve Bayes and perceptron based classification models. More recently Ikeda et al. (2013) has used community mining for classifying tweets. They formulated hybrid text-based and community-based methods to classify tweets based on demographics for a large dataset.

Most of the major works on text features-based gender classification of microblogs are based on word features. Classification based on the word features generally give reasonable accuracy for the dataset used; however, they are heavily dependent on the words used in the text. The classification algorithms used in the above cases may not satisfactorily capture the latent features of tweeting behavior which go beyond the topic being discussed in tweets, and are dependent on the

hidden nuances of the writing style of the genders. These features could be important for correctly predicting the gender of the author of a new tweet, which might be written in a different context. In such cases capturing authorial style becomes indispensable. We have tried to overcome this issue by using features that are mostly independent of the topic of discussion in the tweets. In our method, we focus more on the authorial style of both the genders, which are better captured using function words and part of speech n-grams. Koppel (2002) states that categorization by topic is typically based on keywords which reflects a document's content, whereas categorization by author style uses precisely those features which are independent of context.

Literature states that function words have methodological advantages in the study of authorial style (Binongo 2003). In one of his papers, Kestemont (2014) states the following properties of function words:

1. All authors writing in the same language and over the same period are bound to use the very same function words. Function words are therefore a reliable base for textual comparison.
2. Their high frequency makes them interesting from a quantitative point of view because we have many observations for them.
3. The use of function words is not strongly affected by a text's topic or genre: the use of the article 'the', for instance, is unlikely to be influenced by a text's topic.
4. The use of function words seems less under an author's conscious control during the writing process.

Any similarities between texts with respect to function words are therefore relatively content-independent and can be far more easily associated with authorship than topic-specific stylistics (Kestemont 2014).

Use of part of speech tags has been common for text categorization in regular text and blogs (Argamon et al. 2009; Rao et al. 2010). Literature states that the use of parts-of-speech n-grams is a relatively efficient way to capture the heavier syntactic information, which is useful for distinguishing writing styles (Baayen et al. 1996). Extant literature also states that parts of speech used in a text are mostly independent of the topic under discussion (Koppel 2002; Argamon et al. 2007).

Consequently, one can say that function words and part of speech n-grams are not affected by the topic of discussion in the text and hence are better features to classify text that depend on the author's writing style. We attempt to classify tweets based on these features and their various combinations. We achieve higher accuracy and F-measures for these two feature types as compared to other common features used for classification such as words, character n-grams etc. (Järvelin et al. 2007). We use a much smaller dataset (3000 tweets) than usually used in twitter-based classification and found higher accuracy than the commercially available software (e.g. Gender Guesser, Gender Genie). The results also show comparable accuracy and F-measure to the earlier research on the subject matter.

3 Methodology

A Twitter user profile can provide information about the user's screen name, full name, location, URL and personal description. The user mandatorily provides the screen name, revealing the rest of the information is done at the user's discretion. It's important to emphasize that gender information is not a required field for having a Twitter account. For a small dataset like ours, visiting individual profile to excavate gender information for our training set makes sense. Although the method is labor intensive, it has been extensively used in extant literature (Rao et al. 2010; Miller et al. 2012). It should also be noted that for large datasets, our method might not be suitable and other methods such as—automatically associating blogger profile information to the associated Twitter account can be used (Burger et al. 2011) (Fig. 1).

10,000 Tweets ranging across various women related issues like—abortion, female literacy, violence against women, female empowerment, women rape, gender equality, gender based harassment, forced prostitution of women, domestic violence against women, female infanticide and women health were downloaded. The tweets were cleaned by removing retweets and ambiguous tweets where ever the gender of the tweeter could not be ascertained. The remaining tweets were manually labeled after ascertaining the gender of the author by visiting each individual profile and looking for keywords (mom of two, husband by profession etc.), the profile picture and any other information to confirm the gender of the person. While labeling, we kept one tweet from each user. This reduced the dataset to roughly 3000 users with about 1800 females. To train and test the classifiers, the data was split into two sets randomly. The dataset was randomly divided into a ratio

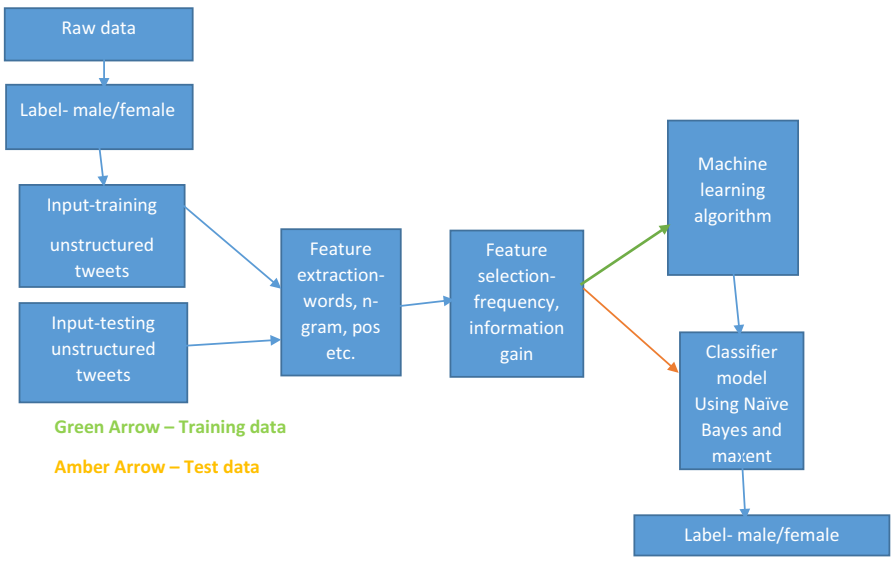

Fig. 1 Gender classification based on supervised learning

of 75–25. The mentioned ratio has been extensively applied in classification literature (Schürer and Muskal 2013). A tenfold cross-validation was performed on the training set. In choosing the training testing ratio the stress is on generalizability of the results, which is achieved by the K-fold cross validation as explained later in this section (Domingos 2012).

One needs to ensure that the training data doesn't over fit the training set as it could drastically distort the result for the test set. This is usually addressed by the K-fold cross validation. For our purposes we use $K = 10$ which is the usual norm in classification data training (Pennacchiotti and Popescu 2011). A tenfold cross

Table 1 Different feature types extracted from twitter datasets

Sl. no	Feature	Definition	Example
1	Content words	Content words typically are a noun, verb, adjective, or adverb, that carries semantic content, bearing reference to the world independently of its use within a particular sentence (Winkler 2012)	School, beer, run, black, teach
2	Function words	Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker (Klammer et al. 2000)	The, These, in, can, my
3	Part of speech tags	It is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context (Church 1989)	This/PNN, is/VB, a/ART dog/NN
4	Part of speech n-grams	An n-gram model is a type of probabilistic language model for predicting the next item in a sequence in the form of a $(n - 1)$ order Markov model. The prediction could be done on the basis of a single preceding item (unigram), two preceding items (bigram) or more items (trigram, four gram etc.). In our case the items are part of speech of the words used in the sentences (Koppel 2002)	PNN,VB,ART,NN
5	Character n-grams	These are similar to other n-grams like word and part of speech. Here the items are letters or characters used in the words of a sentence (Järvelin et al. 2007)	'a', 'b', 'o', 'v', 'e'
6	Function words + part of speech n-grams	Combined function words and part of speech n-grams and used them as a single feature for classification	This, VB, in, NN, ART
7	All words	all the words present in the tweets including the stop words	This, is, a, dog
8	Content words + function words + part of speech n-grams	combination of the most informative content words, the function words and the part of speech n-grams as features	School, this, VB, in, Beer, ART

6,7,8 are combinations of 1,2,3,4

validation entails dividing the dataset into ten equal random folds and nine of them are used for training and one for testing or validation. The whole process is repeated ten times with each of the sub folds being used for validation exactly once. This ensures that the model generalizes to an independent dataset and doesn't over-fit (Kohavi 1995).

Usable features from tweets were extracted and selected from the training set. The features were then tested for accuracy and F-measure on the test set. We started with a small number of tweets and progressively increased the number to observe its effect on the classification accuracy and the F-measure. One must bear in mind that, for a small dataset, the method for manually cleaning and labeling tweets is standard in supervised learning. We have emphasized extracting features that have not been used in extant literature.

We now explain the feature extraction and feature selection methods used in our work.

3.1 Feature extraction

Feature extraction is a method used to reduce the amount of resources required to describe a large dataset (Guyon and Elisseeff 2003). When analyzing complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power. It may also lead to the formation of a classification algorithm, which over fits the training sample and generalizes poorly to new samples. Hence, feature extraction becomes essential while dealing with classification problems with large number of variables.

We have extracted a comprehensive list of linguistic features for our classification job. Using different features let us compare the results across the features as listed in the Table 1.

3.2 Feature selection

Feature selection is a process through which a subset of relevant features is selected for model formation (Guyon and Elisseeff 2003). This removes the redundant and/or irrelevant and/or less important features. Though it causes some loss of information, the objective of feature selection is to consider only the most relevant features with minimum loss of information. It is a well-accepted and almost a mandatory method in text classification of any kind (Mukherjee and Liu 2010; Rao et al. 2010).

We have applied two feature selection criteria in our models as discussed below.

3.3 Information gain

We use information gain as one of the feature selection criteria in our model. This technique measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document.. Information gain is often employed as a term-goodness criterion in the field of machine learning (Lee and Lee 2006).

Information gain is an entropy based metric (Zhang and Zhang 2010). It can be measured by the formula shown below:

$$\mathrm{IG(f)} = -\sum_{c,c} \mathrm{P(c)} \log \mathrm{P(c)} + \sum_{f,f} \mathrm{P(f)} \sum_{c,c} \mathrm{P(c/f)} \log \mathrm{P(c/f)}$$

Here, "IG (f)" is the information gain for the given class. "C" denotes the classes {male, female} and "f" = $\{f_1, f_2, \ldots, f_n\}$ are the set of features. The objective of information gain is to consider only the most relevant features with minimum loss of information.

3.4 Term frequency

Another feature selection method is frequency-based feature selection, which is, selecting the terms that are most common in the class. Frequency can be either defined as document frequency (the number of documents in the class that contain the term) or as collection frequency (the number of tokens of that occur in documents) (Azam and Yao 2012). The basic assumption is that, the rare terms in the classes are either non-informative for category prediction or not useful in global performance (Yang and Pedersen 1997).

4 Classification method

Broadly there are two classification models in NLP—generative and discriminative. Generative classifiers learn the joint probability of the inputs and the labels (male/female, in our case), and make the prediction by using the Bayes' rule to select the most likely label. The discriminative classifiers model the posterior probability directly or learn a direct map of inputs to the class label (Jordan and Ng 2002). Researchers have used both types of classifiers in the past. Yan and Yan (2006) used a generative classifier (Naïve Bayes) and Rao et al. (2010) used a discriminative classifier (SVM). We have formulated both the types of classification models, the Naïve Bayes model (generative classifier) and the maximum entropy model (discriminative classifier).

4.1 Naïve Bayesian classifier

The Naïve Bayes classifier is a popular classification algorithm used extensively in document classification (Yan and Yan 2006; Argamon et al. 2007; Mukherjee and Liu 2010). We have shown how the Naïve Bayes classifier works in the case of text classification. We considered a document vector model (Manning and Schutze 1999; Weikum 2002) for representing a document with the help of terms that can be used as inputs.

Let's consider a tweet with some features of our interest T = $(F_1, F_2, \ldots, F_n)$. Here F can be any of the features based on which we would like to classify the tweets (content words, n-gram part of speech tags, function words etc.). Given the

tweet "T" we would like to predict whether it belongs to a particular gender, viz. male or female.

Using Bayes' theorem we can write,

$$p(C/F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n/C)}{p(F_1, \ldots, F_n)} \quad (1)$$

where, C = {Male, Female}. F_i represents the features selected as inputs for developing the classification model as per Table 1.

The Naïve Bayes assumption for a classification task is as shown below:

$$p(F_1, ..., F_n|C) = p(F_1|C)p(F_2|C), \ldots, p(F_n|C) \quad (2)$$

The assumption of independence between or amongst the features is considered in the above expression. In the case of tweets it will mean that the two words (a feature) in a tweet occur independent of each other. Although the assumption is simplistic it has been shown to work well in earlier research (Yan and Yan 2006).

This equation could now be written as,

$$p(C/T) = \frac{p(C)p(F_1|C)p(F_2|C), \ldots, p(F_n|C)}{p(T)} \quad (3)$$

We then compute the ratios of the posterior probabilities P(C = male|T) and P(C = female|T) of the two classes for a given document. This is done by calculating the prior probabilities p(C) and the conditional probabilities of P(Fi|T). The tweet is then classified to the class that yields the higher probability.

4.2 Maximum entropy classifier

Unlike the Naïve Bayes classifier, the maximum entropy classifier doesn't assume that the features are conditionally independent of each other. Maximum entropy is therefore a less restrictive model than the Naïve Bayesian model (Juan et al. 2007). It is based on the principle of maximum entropy and from all the models which fit the training data, it selects the one which has the highest entropy. The maximum entropy classifier requires more time to train compared to Naïve Bayes due to the optimization problem that needs to be solved in order to estimate the parameters of the model.

We construct a stochastic model (Berger et al. 1996) that accurately represents the behavior of the process. We take as input the contextual information "a" (function words, unigram, bigram etc.) of a document and produce the output value "b".

The initial step of constructing this model is to collect training data which consists of samples represented in the following format: (a_i,b_i) where the a_i includes the contextual information of the document and b_i its class. The next step is to summarize the training sample in terms of its probability distribution:

$$p(a, b) = \frac{1}{N} \times \text{number of times that}(a, b)\text{occurs in the sample set} \quad (4)$$

where N is the size of the training set.

We use the above empirical probability distribution in order to construct the statistical model of the random process that assigns texts to a particular class by taking into account their contextual information.

We use the following function:

$$f_j(a,b) = \begin{cases} 1 & \text{if } b = \text{Li and a contains Ki} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where f_j is the feature function that returns 1 when the class of the function is L_i and the document contains the word K_i. We express any statistic of the training dataset as the expected value of the appropriate binary-valued indicator function f_j.

The expected value of f_j with respect to the distribution p(a, b) is:

$$p(f_j) \equiv \sum_{a,b} p(a,b) f_j(a,b) \tag{6}$$

If each training sample (a, b) occurs once in training dataset then p(a, b) is equal to 1/N.

We constrain the expected value that the model assigns to the expected value of the feature function fj. The expected value of feature fj with respect to the model $p\left(\frac{b}{a}\right)$ is equal to:

$$p(f_j) \equiv \sum_{a,b} p(a) p\left(\frac{b}{a}\right) f_j(a,b) \tag{7}$$

where p(a)is the empirical distribution of a in the training dataset and it is usually set equal to 1/N.

By constraining the expected value to be the equal to the empirical value:

$$\sum_{a,b} p(a,b) f_j(a,b) = \sum_{a,b} p(a) p\left(\frac{b}{a}\right) f_j(a,b) \tag{8}$$

Equation (8) is the constrain equation which depends on the number of feature functions.

The constraint in Eq. (8) can be satisfied by multiple models, however according to the principle of maximum entropy the model should be the most uniform amongst the ones which satisfy the constraint. One could also say that the model should have the maximum entropy to be selected:

$$p_{max} = \arg\max_{p \in L} \left(-\sum_{a,b} p(a) p\left(\frac{b}{a}\right) \log p\left(\frac{b}{a}\right) \right) \tag{9}$$

It now becomes an optimization problem with Eq. (8) as the constraint.

Both the above mentioned techniques address the classification problem considering the classification boundary to be linearly separable. This gives

satisfactory result in our case. The research could be extended to nonlinear classifiers like K-nearest neighbors and support vector machine (Rao et al. 2010).

5 Results and discussion

We now report the gender classification ability of the classification algorithms based on the various feature types. The performance of these systems was measured by a variety of metrics such as precision, recall, accuracy and F-measure. Accuracy is the percentage of instances predicted in the correct classes in a classification problem. However, in case of unbalanced classes, accuracy can give spurious results, in such cases F-measure in classification is a better metric. It is a measure that combines precision and recall by calculating the harmonic mean of precision and recall.

It is denoted in its common form by the following formula:-

$$\text{F - Measure} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

where precision is the number of retrieved instances that are relevant and recall is the number of relevant instances which have been retrieved.

The following confusion matrix illustrates:

Confusion matrix	(Predicted class) Yes	No
(Actual class) Yes	TP	FN
No	FP	TN

$$\text{Here, precision} = \text{TP}/(\text{TP} + \text{FP}), \quad \text{recall} = \text{TP}/(\text{TP} + \text{FN}), \quad \text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

where, TP = true positive, TN = true negative, FP = false positive, FN = false negative.

F-measure has been used in previous studies in gender identification as an overall assessment of performance of a classifier as it takes into account both precision and recall (Pennacchiotti and Popescu 2011). When measured by these metrics, each algorithm demonstrates its gender prediction capability. Once the data is trained on the training data set, both the maximum entropy algorithm and the Naïve Bayes algorithms are run on the test set. We used several tweet datasets with increasing number of tweets in each of them to identify the best feature type. We found that "Part of speech n-grams", "function words" and "Part of speech n-grams and function words" give the best results for both the measurement metrics across the datasets. This is in line with our initial claim that features which are independent of text could be better at classifying tweets than the other more popular features.

The following Tables 2 and 3 summarize the accuracy, F-measure and the best feature type for each of the tweet datasets. The best feature types across datasets in Tables 2 and 3 are marked in bold.

Table 2 Best feature type across the datasets (accuracy)

Dataset	Best feature type	Accuracy
500 Tweets	ngrampos + funcW	0.6
1000 Tweets	ngrampos + funcW	0.63
1500 Tweets	ngrampos	0.63
1856 Tweets	ngrampos	0.66
2325 Tweets	ngrampos	0.67
3000 Tweets	**ngrampos**	**0.71**

ngrampos part of speech n-grams, *funcW* function word

Table 3 Best feature type across the datasets (F-measure)

Dataset	Best feature type	F-measure
500 Tweets	funcW	0.72
1000 Tweets	funcW	0.79
1500 Tweets	funcW	0.89
1856 Tweets	**ngrampos + funcW + contW**	**0.95**
2325 Tweets	ngrampos	0.77
3000 Tweets	ngrampos	0.75

ngrampos part of speech n-grams, *funcW* function word, *contW* content word

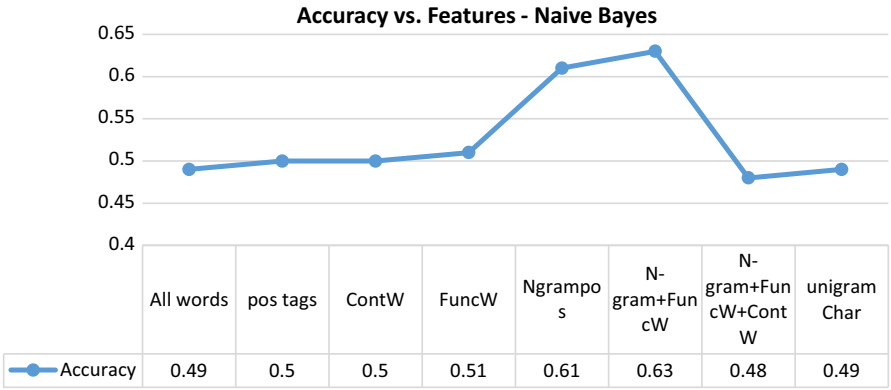

Fig. 2 Accuracy versus feature type, Naïve Bayes (1000 tweets)—test data. *pos tags* part of speech tags, *ContW* content word, *FuncW* function words, *Ngrampos* part of speech n-grams, *unigram char* character unigram, test data

The Figs. 2 and 3 compare the different feature types on accuracy and F-measure across the 6 twitter datasets we used for our research. The two classifiers used in our work have been compared on accuracy and F-measure across various feature types. We found that the Naïve Bayesian classifier performs better than the maximum entropy classifier on most occasions.

We have shown the variation in accuracy across different feature types only for the Naïve Bayes classifier for this dataset, as the maximum entropy classifier gave almost similar results and the two graphs could not be substantially differentiated.

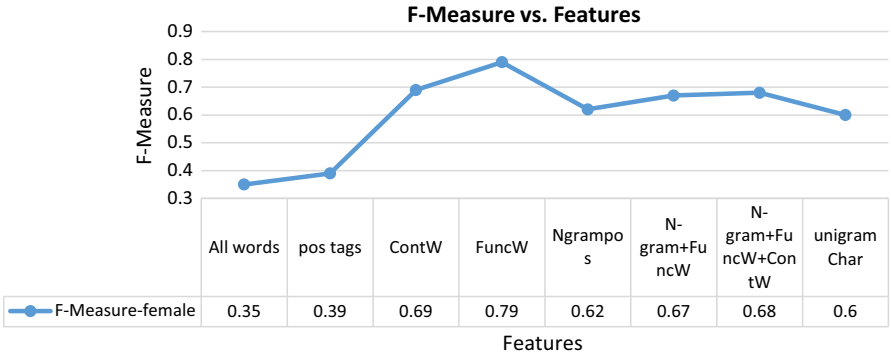

Fig. 3 F-measure versus feature type (1000 tweets)—test data

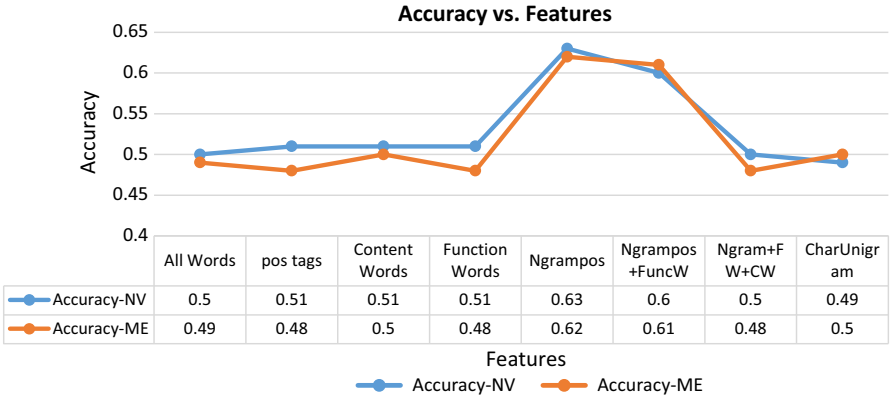

Fig. 4 Accuracy versus features—both classifiers (1500 tweets)—test data. *NV* Naïve Bayes, *ME* maximum entropy

However, we do get different results for both the classifiers for the other datasets as can be observed in the Figs. 4, 5, 6, 7, 8, 9, 10 and 11.

In the Fig. 11, we observe that for the authoritative dataset of 3000 tweets, part of speech n-grams give the best performance across both the classifiers. From our experiments, part of speech n-grams come out to be the best features for classifying the gender of microblog authors. It's worth noting that part of speech n-grams are authorial style based features.

In the Figs. 12 and 13 we show accuracy and F-measure plotted against the tweet datasets we have used. On the X-axis we have the tweet datasets with progressively increasing number of tweets and on the corresponding Y-axis, we have accuracy and F-measure.

It can be observed that with the increase in the size of the datasets the accuracy improves almost linearly. It can be inferred that for larger datasets the accuracy would improve further. However, as already mentioned before, it has to be ensured

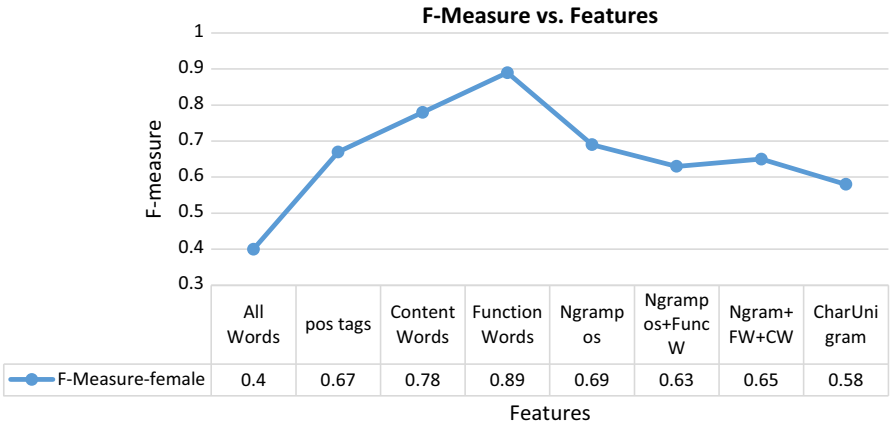

Fig. 5 F-measure versus feature types (1500 tweets)—test data

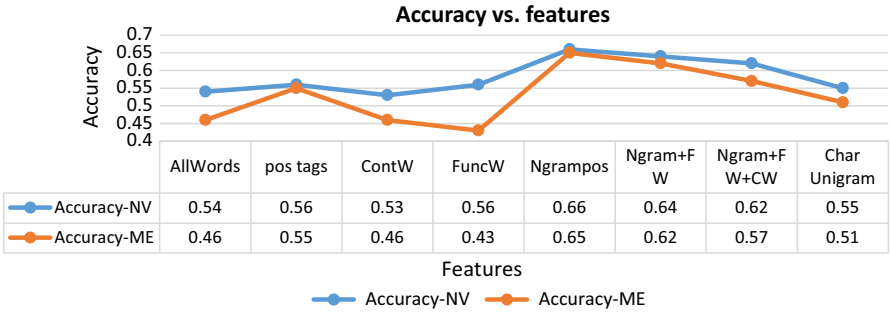

Fig. 6 Accuracy versus features—both classifiers (1856 tweets)—test data. *NV* Naïve Bayes, *ME* maximum entropy, test data

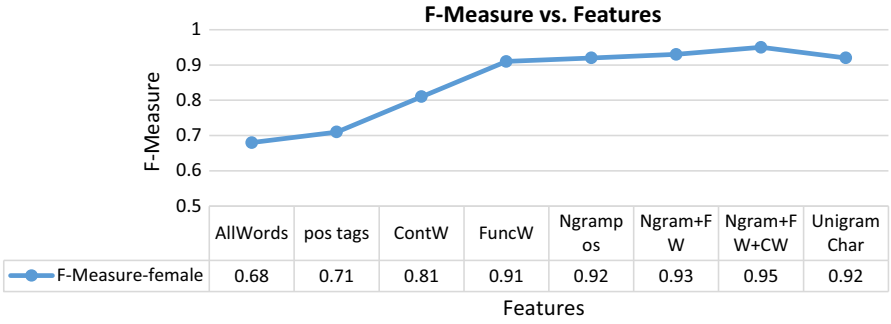

Fig. 7 F-measure versus feature types (1856 tweets)—test data

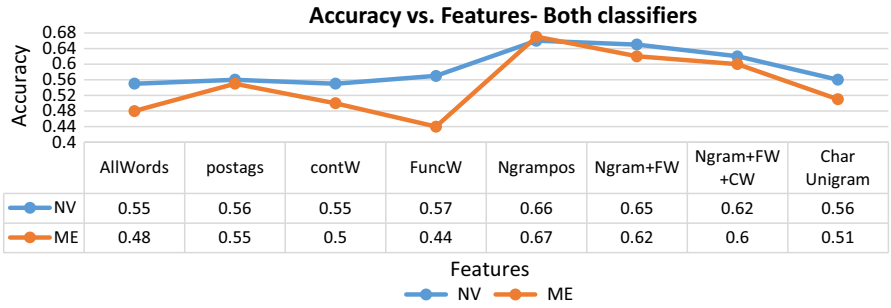

Fig. 8 Accuracy versus features—both classifiers (2325 tweets)—test data. *NV* Naïve Bayes, *ME* maximum entropy, test data

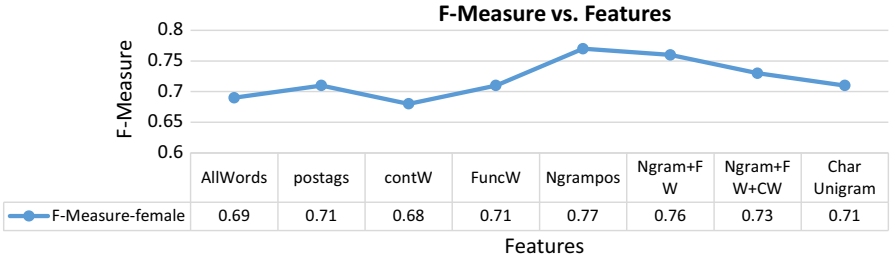

Fig. 9 F-measure versus feature types (2325 tweets)—test data

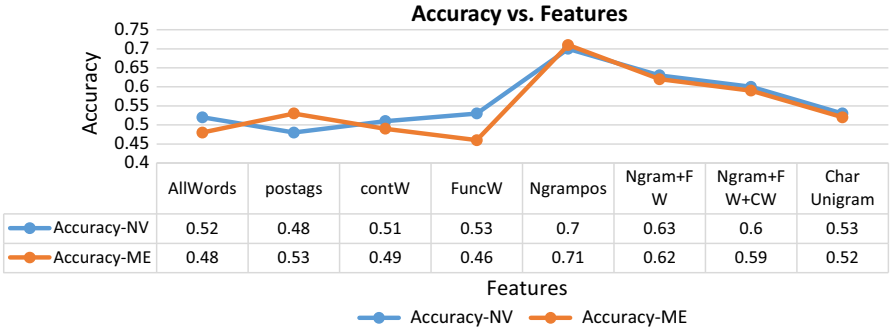

Fig. 10 Accuracy versus features – both classifiers (3000 tweets)—test data. *NV* Naïve Bayes, *ME* maximum entropy

that the gain in accuracy of the classification is not at the cost of substantial loss in the precision and recall.

From Fig. 13 we find that the Dataset 4 with 1856 tweets gives the highest F-measure among all the datasets. It can be observed in Fig. 13 that the F-measure drops for the last two datasets which is unusual for an increasing dataset size. As stated earlier in this section, F-measure is the harmonic mean of precision and

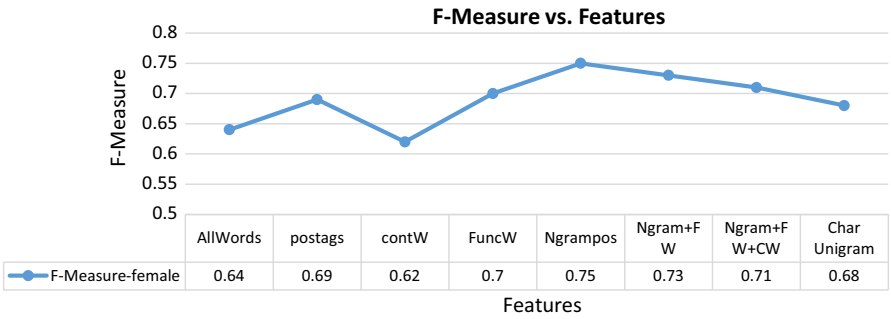

Fig. 11 F-measure versus feature types (3000 tweets)—test data

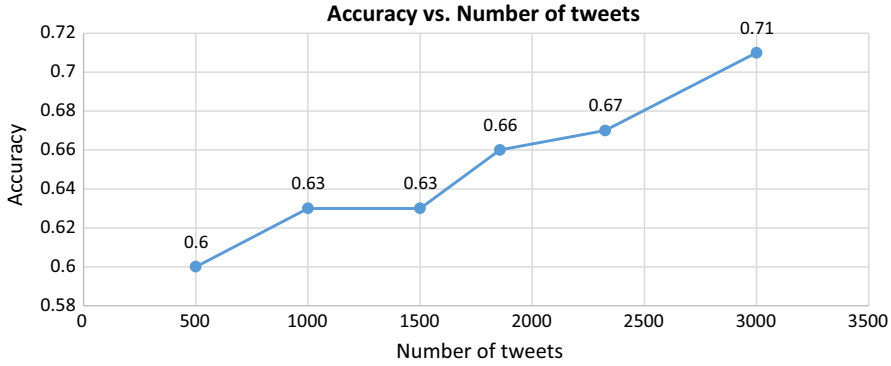

Fig. 12 Accuracy trend across tweet datasets—test data

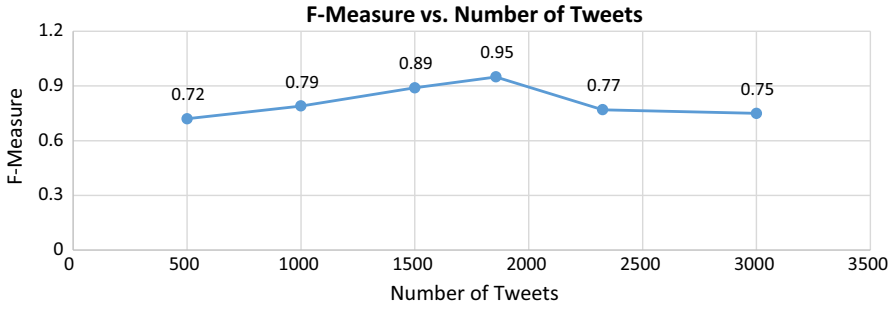

Fig. 13 F-measure trend across tweet datasets—test data. F-measure trend across tweet datasets, test data

recall. It balances the overall result by compensating for any extreme value. In case one of the two parameters gets exceptionally high or low, the resultant F-measure is affected as is observed here. This has also been observed in extant literature. In Yan and Yan (2006), there is no considerable increase in the F-measure for the last two

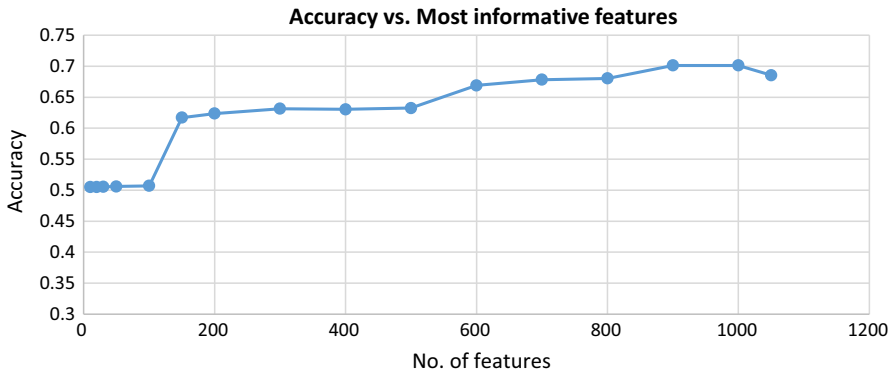

Fig. 14 Accuracy versus most informative features—test data. Accuracy versus most informative features, test data

datasets, however for the first four datasets there is a steep increase. Even in the case of Miller et al. (2012) the increase in the F-measure for the last four tweet lengths is miniscule. In Pennachiotti and Popescu (2011) for the Starbucks fans task a fall in F-measure can be observed while considering more features in social and linguistic features.

There is no effect on the accuracy, which continues to improve as it measures correctly classified instances across all the classes. It should be noted that in spite of falling in the overall value, the F-measure is still above 75 % which is a good performance.

In our Fig. 14, we tried to study the variation in the accuracy of classification by plotting accuracy against number of features. Increase in accuracy can be observed with increasing number of features but at a diminishing rate.

One interesting observation Fig. 14 offers is that most important 100 features give us an accuracy of 63 % whereas when we increase the number of features to 1000 the accuracy improves to 71 %. There is only about 8 % loss in accuracy with 900 lesser features. This emphasizes the importance of feature selection as a method for reducing features in machine learning. It can be inferred that even a small set of features could give reasonably good results. This becomes critical when the training set is large.

We have also identified the most informative function words and part of speech n-grams, which help in differentiating between the genders based on the likelihood of the feature type to appear in a class. For example the word ‘never’ is 5.7 times more likely to appear in the class male than class female.

5.1 Best differentiators: function words

The Table 4 lists the best function word differentiators between the two classes. The top function words used by men in our dataset are “never, either, couldn’t, though and six”. The top function words women use in our dataset are “today, done, gets, and us”. It’s not so easy to make any direct inference on the basis of the function

Table 4 Best function word differentiators Invalid source specified

Word	Gender	Likelihood
Never	Male	5.7:1.0
Today	Female	3.9:1.0
Either	Male	3.8:1.0
Couldn't	Male	3.8:1.0
Though	Male	3.8:1.0
Six	Male	3.0:1.0
Another	Male	2.7:1.0
Done	Female	2.6:1.0
Gets	Female	2.3:1.0
Us	Female	2.3:1.0

words used by both the genders. But on a closer look some patterns emerge, like, women appear to be more affirmative in their communication on Twitter as compared to men. If applied to a larger dataset, more prominent observations could be made which might be crucial in identifying hidden patterns similar to the one observed here. The results obtained here are indicative and not conclusive. Next, we move on to identify the part of speech n-gram features for both the genders.

5.2 Best differentiators: part of speech n-gram

The Table 5 lists the most informative part of speech n-gram features that differentiate between the two classes. The most informative features column provides the acronym for the part of speech n-gram used by either gender. The description column gives the definition and a brief description of the most informative features. The gender and likelihood columns elucidate the gender/class, which uses the feature most and the likelihood of the feature to appear in that class respectively.

It can be observed from the Table 5 that the male differentiators based on part of speech n-grams are personal pronouns and verbs like "to have "etc., past participle like "to be" and third person singular and verb together, like—it's, he's, she's. Female differentiators based on the same feature are noun, singular reflexive pronoun and the verb "to be" in the present tense. Clearly, a difference in the use of part of speech applied through n-grams can be observed between the two genders. For example—women use nouns more than men in tweets and are 11.9 times more likely to use it when compared to men, likewise men use more of the verb "to be" in the past participle form then women do and are 3.2 times more likely to use it than women.

5.3 Comparison with other classifiers

In the Table 6 we have compared the accuracy of our algorithm with the accuracy of two commercially available gender classification software—Gender Genie and Gender Guesser. It can be observed that our classification method outperforms both

Table 5 Part of speech n-gram best differentiating features

Most informative features	Gender	Likelihood	Description
NNS$-TL	Female	11.9:1.0	noun, plural, common, genitive taxpayers' children's members' states' women's cutters' motorists' steelmakers' hours' nations' lawyers' prisoners' architects' tourists' employers' secretaries' rogues'
BEM	Female	6.8:1.0	verb "to be", present tense, 1st person singular am
NN	Female	4.4:1.0	noun, singular, common failure burden court fire appointment awarding compensation mayor interim committee fact effect airport management surveillance jail doctor intern extern night weekend duty legislation tax office
PPSS + HVD	Male	3.8:.0	pronoun, personal, nominative, not 3rd person singular + verb "to have", past tense I'd you'd we'd they'd
NNS$	Female	3.5:1.0	noun, plural, common, genitive taxpayers' children's members' states' women's cutters' motorists' steelmakers' hours' nations' lawyers' prisoners' architects' tourists' employers' secretaries' rogues'
BEN	Male	3.2:1.0	verb "to be", past participle been
PPL	Female	3.0:1.0	pronoun, singular, reflexive itself himself myself yourself herself oneself ownself
PPS + HVZ	Male	2.8:1.0	pronoun, personal, nominative, 3rd person singular + verb "to have", present tense, 3rd person singular it's he's she's

Table 6 Comparison with other classifiers

System	Accuracy
Gender Genie	61.69
Gender Guesser	63.78
Our method	**71.00**

of these software. We have shown an improvement of over 7 % in classification accuracy with our method over the software. We have marked the accuracy obtained through our method in bold.

6 Conclusion

Anonymity of the user on the Internet is a common occurrence. This makes the texts from the user a useful source for extracting relevant information about him/her. One of the key findings which can be utilized in a number of areas if obtained with reasonable accuracy is the gender of the user. Knowing the gender of an individual could help in product recommendations specific to the user's requirement, understanding gender opinions on social issues, and can aid in detecting cybercrimes (by ascertaining the gender of the suspect). Gender classification of

online unstructured text data is a relevant business problem. We have tried to classify the tweets in our dataset by extracting features which best capture authorial style. This has been an effective way to classify gender in case of regular text (Argamon et al. 2003, 2007) but unapplied in case of microblogs due to complexity in capturing such features from limited text.

The data available from social networking microblog sites such as Twitter is unstructured and often restricted by a maximum length constraint which makes it difficult to use them for any classification job. In this paper, we have used data from Twitter and extracted novel features, like—function words and Part of speech n-grams. We also extracted other commonly used features, like- all words in text, content words and character n-grams. Further, we applied two feature selection methods namely—term frequency and information gain to reduce the number of features extracted to only the most relevant ones. The relevant features extracted and selected were then classified using Naïve Bayes and maximum entropy algorithms. The algorithm performances were compared based on accuracy and F-measure. We found that the feature "part of speech n-gram" gave better classification accuracy and F-measure than the other features, like—words and character n-grams across the datasets. Optimal results reveal that part of speech n-grams are the best features for classifying the gender of microblog authors. Naive Bayesian and maximum entropy classifiers have similar precision, recall and accuracy performance with this feature. This establishes that authorial style based features can be applied to distinguish between the genders based on their writing behavior on microblogs like- Twitter, and Facebook and are more universal in nature as compared to other features.

In future, the research could be extended by considering other feature selection techniques like IDF, TFIDF etc. and capturing their effect on the overall result. Also, other classification techniques which consider a non-linear decision boundary, like—SVM, neural networks or a Bayesian network could be applied. The research could also be extended to other related areas which have been traditionally difficult to classify such as detection of sarcasm in unstructured text.

References

Alowibdi JS, Buy UA, Yu P (2013) Language independent gender classification on Twitter. In: Proceedings of 2013 IEEE/ACM international conference on Advances in social networks analysis and mining (ASONAM). IEEE, Niagara Falls, pp 739–743. doi:10.1145/2492517.2492632

Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. Text Interdiscip J Study Discourse 23:321–346. doi:10.1515/text.2003.014

Argamon S, Koppel M, Pennebaker J, Schler J (2009) Automatically profiling the author of an anonymous text. Commun ACM. doi:10.1145/1461928.1461959

Argamon S, Koppel M, Pennebaker JW, Schler J (2007) Mining the blogosphere: age, gender and the varieties of self-expression. First Monday 12(9). doi:10.5210/fm.v12i9.2003

Azam N, Yao J (2012) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst Appl 39:4760–4768. doi:10.1016/j.eswa.2011.09.160

Baayen H, Van Halteren H, Tweedie F (1996) Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Lit Linguist Comput 11:121–132

Berger A, Pietra V, Pietra S (1996) A maximum entropy approach to natural language processing. Comput Linguist 22:39–71. doi:10.3115/1075812.1075844

Binongo JNG (2003) Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance 16:9–17. doi:10.1080/09332480.2003.10554843
Burger JD, Henderson J, Kim G, Zarrella G (2011) Discriminating gender on Twitter. Test 146:1301–1309. doi:10.1007/s00256-005-0933-8
Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55:78. doi:10.1145/2347736.2347755
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. doi:10.1023/A:1012487302797
Hota SR, Argamon S, Koppel M, Zigdon I (2006) Performing gender: automatic stylistic analysis of shakespeare's characters. Digit Humanit 1:82–88
Ikeda K, Hattori G, Ono C et al (2013) Twitter user profiling based on text and community mining for market analysis. Knowl-Based Syst 51:35–47. doi:10.1016/j.knosys.2013.06.020
Järvelin A, Järvelin A, Järvelin K (2007) S-grams: defining generalized n-grams for information retrieval. Inf Process Manage 43:1005–1019. doi:10.1016/j.ipm.2006.09.016
Jordan MI, Ng AY (2002) On disriminative vs. generative classiers: a comparison of logistic regression and naive Bayes. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 841–848
Juan A, Vilar Torres D, Ney H (2007) Bridging the gap between naive Bayes and maximum entropy text classification. In: Proceedings of the 7th international workshop on pattern recognition in information systems (PRIS). INSTICC Press, Setúbal, pp 59–65
Kestemont M (2014) Function words in authorship attribution from black magic to theory? In: 3rd Workshop on computational linguistic for literature (CLfL 2014), pp 59–66
Klammer T, Schulz M, Della Volpe A (2000) Analyzing English grammar, 6th edn. Pearson Education
Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence
Koppel M (2002) Automatically categorizing written texts by author gender. Lit Linguist Comput 17:401–412. doi:10.1093/llc/17.4.401
Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manage 42:155–165. doi:10.1016/j.ipm.2004.08.006
Manning CD, Schutze H (1999) Foundations of statistical natural language processing. MIT press, Cambridge
Miller Z, Dickinson B, Hu W (2012) Gender prediction on twitter using stream algorithms with n-gram character features. Int J Intell Sci 02:143–148. doi:10.4236/ijis.2012.224019
Mukherjee A, Liu B (2010) Improving gender classification of blog authors. In: Proceeding EMNLP '10 proceedings of the 2010 conference on empirical methods in natural language processing, pp 207–217
Peersman C, Daelemans W, Van Vaerenbergh L (2011) Predicting age and gender in online social networks. In: International conference on information and knowledge management proceedings, pp 37–44. doi:10.1145/2065023.2065035
Pennacchiotti M, Popescu A-M (2011) A machine learning approach to Twitter user classification. ICWSM 11:281–288
Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop search mining user-generated contents—SMUC'10, p 37. doi:10.1145/1871985.1871993
Schürer SC, Muskal SM (2013) Kinome-wide activity modeling from diverse public high-quality data sets. J Chem Inf Model 53:27–38. doi:10.1021/ci300403k
Weikum G (2002) Foundations of statistical natural language processing. ACM SIGMOD Rec 31:37. doi:10.1145/601858.601867
Winkler E (2012) A basic course in linguistics. Bloomsbury Publishing, London
Yan X, Yan L (2006) Gender classification of weblog authors. In: AAAI spring symposium series on computational approaches to analysing weblogs, pp 228–230
Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. Mach Learn Work Then Conf. doi:10.1093/bioinformatics/bth267
Zhang C, Zhang P (2010) Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA