Semantic-Based Gender Identifications through User-Generated Contents

Mohammed Ali Ibrahim Eltaher

Faculty of Information Technology, University of Tripoli, Tripoli, Libya
m.eltaher@uot.edu.ly

Abstract. User gender is crucial information for personalized services and applications in online social networks. It impacts areas such as recommendation systems, advertising, and connection discovery. However, user gender information may be hidden or not specified in online social networks, leading to inaccuracies or limitations in various applications. The daily interactions of billions of users on online social networks like Flickr contribute to creating vast amounts of user-generated content. This content includes multiple media such as images, videos, and textual information. The primary aim of this paper is to address the challenge of identifying the gender of users. Our approach involves a semantic-based data technique. Using a semi-automatic image tagging system implies a process where images are labeled or categorized with automation, potentially improving efficiency and accuracy. We employ two classification algorithms for gender identification: Naive Bayes and Support Vector Machines (SVM), where data are typically represented as feature vectors. Our experimental results on more than 149,700 Flickr users demonstrate an accuracy of over 84% for gender identification. This suggests that combining Naive Bayes and SVM algorithms, with data represented as feature vectors, has proven effective in classifying gender based on user-generated content.

Keywords: Gender identification, user profiling, semantic content, and user-generated content.

1 Introduction

A user profile is a key element of information systems. It has been instrumental across various fields including healthcare, banking, social media, e-commerce, security, access control, and social networking [1]. Indeed, the ability to predict user demographics, such as gender, in the context of social media has significant implications for various applications and services. The extraction of such information can be leveraged for targeted marketing, personalized recommendations, and improved user experiences. Moreover, understanding the gender distribution of users allows marketers to tailor their strategies and advertisements more effectively. Targeted marketing based on gender can increase the relevance of ads, leading to higher engagement and conversion rates. Gender prediction can contribute to content filtering and moderation, ensuring that users are exposed to content that aligns with their preferences and demographics.

With a growing amount of user-generated content, efficient techniques for discovering patterns become essential. This can include identifying trends, preferences, or common themes among the tagged images. For example, Flickr has become more popular by allowing people to easily upload, share, and annotate multimedia objects with keywords. Labeling the multimedia objects, i.e., images, with a set of keywords is known as image tagging. Most of the social user mining tasks depend on the availability and quality of the tagging system. However, the existing studies show that tags are impressive, ambiguous, and overly personalized. By incorporating semantic concepts, the proposed approach aims to enhance the quality of tagging. Semantic concepts can add a layer of meaning and context to tags, potentially reducing ambiguity and improving relevance. To overcome the above problems with tags, we are proposing a solution involving a semi-automatic image tagging system called " akiwi [1] " that incorporates semantic concepts.

In this paper, we propose a novel approach for gender identification through user-generated content on Flickr, utilizing images and employing a semantic-based information technique. Our approach is different from previous methods by relying on a semi-automatic image tagging system. This suggests a combination of automated processes and human input for tagging images, with a focus on semantic understanding. For the classifier, we use a Naive Bayes algorithm with multinomial distributed data, where the integrated data are typically represented as feature vectors as well as SVM. The dataset is sourced from Flickr.com and consists of 149,700 user profiles. Each user profile includes up to 60 photos. Figure 1 illustrates the different stages and components of our methodology.

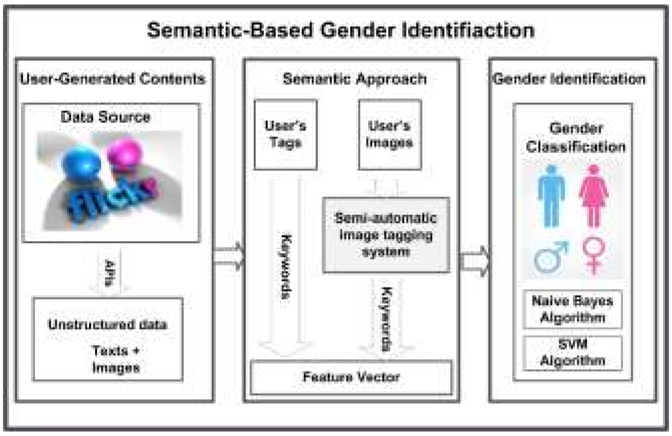

Fig. 1. The different stages and components of our methodology

[1] http://www.akiwi.eu/

2 Related Work

We describe relevant related work in two areas, gender classification and semantic-based contents.

2.1 Gender identification

Certainly, the study of mining demographics, including factors such as gender, ethnicity, age, and marital status, is a rich area of research with implications across various fields. When it comes to user gender specifically, researchers often explore how it impacts user behavior, preferences, and experiences in different domains, such as online platforms, social media, and technology usage. Peersman et al. [2] focus on using a text categorization approach to predict age and gender based on a corpus of chat texts from the Belgian social networking site Netlog [2]. The text type under consideration is described as difficult. This suggests that chat texts may have unique characteristics or challenges that need to be addressed in the prediction task.
Moreover, Burger et al. [3] discussed the development and evaluation of language-independent classifiers for predicting the gender of Twitter users using the content of the tweet text as well as three fields from the Twitter user profile: full name, screen name, and description. In addition, [4] described a research task related to predicting the gender of YouTube users based on two different information sources: comments and the social environment derived from the affiliation graph of users and videos. The key finding is that gender information can be accurately predicted from the social environment. The authors in [5] investigated the relationship between shared images and user gender on Fotolog and Flickr. The study involved a substantial dataset of 3,152,344 images from 7,450 users on these two image-oriented social networks. The key findings indicate that users who share visually similar images are more likely to have the same gender. This observation suggests a potential correlation between the content of shared images and user gender. Miura et al. [6] discussed a method for estimating user gender based on geographical information from social media posts. The proposed method relies on the idea that people with certain attributes tend to visit specific areas, and these areas may vary depending on the attributes. The approach involves creating feature vectors based on geographical information associated with social media sites.

2.2 Semantic-based contents.

Extracting and analyzing meaning-related information from social network data has garnered significant interest from researchers in fields such as natural language processing, corpus linguistics, information retrieval, and data science. A key element of this automated information extraction and analysis is using semantic tagging tools to annotate multimedia data. Various tools have been specifically designed to perform different levels of semantic analysis, including named entity recognition and disambiguation, sentiment analysis, word sense disambiguation, content analysis, and semantic

[2] http://www.netlog.com

role labeling. A common requirement for these tasks, particularly in supervised settings, is the presence of a manually annotated corpus that serves as a knowledge base for training and testing word and phrase-level sense annotations.

The authors in [7] introduces a comprehensive benchmark corpus and methodologies for the semantic tagging task in the Urdu language. The corpus consists of 8,000 tokens, evenly distributed across four domains: news, social media, Wikipedia, and historical texts, with each domain containing 2,000 tokens. It has been manually annotated with 21 major semantic fields and 232 sub-fields using the USAS (UCREL Semantic Analysis System) taxonomy, enabling detailed coarse-grained annotation. Each word in the corpus is tagged with one to nine semantic field labels, facilitating an in-depth semantic analysis of the data. This approach allows them to frame semantic tagging as a supervised multi-target classification problem. To illustrate the utility of our corpus for developing and evaluating Urdu semantic tagging techniques, they extracted local, topical, and semantic features, applying seven different supervised multi-target classifiers.

Indeed, a semi-automatic tagging process can significantly enhance the efficiency and accuracy of tagging multimedia objects, leading to improved quality in tagging and a more effective social user mining process. The authors in [8] represent a step forward to rethink person re-identification via semantic-based pretraining. They propose a straightforward semantic-based pretraining approach to replace the conventional ImageNet pretraining. This method enables the learning of visual representations from textual annotations in downstream re-identification tasks. This underscores the potential of semantic-based pretraining for future investigations.

In the context of automated photo tagging on Flickr's images, the authors in [9], and [10] designed a distance metric that captures the similarity or dissimilarity between images based on their content, making it more effective for tasks like image retrieval and tagging. Wu et al. [9] presented a probabilistic distance metric learning technique (PDML). First, they discover probabilistic side information from the data using a graphical model approach and then present an effective probabilistic RCA algorithm to find an optimal metric from the probabilistic side information. On the other hand, [10] proposed a unified distance metric learning (UDML) method that addresses the challenge of metric learning by leveraging implicit side information present in massive social images on the web. The use of both textual and visual content in a unified learning framework is a noteworthy aspect of their approach. To better understand the details and potential impact of the methods proposed in [9], and [10], it would be helpful to know more about the specific techniques or algorithms used, how the textual and visual information is integrated, and the performance evaluation metrics or benchmarks used to assess the effectiveness of your approach.

Associating visual features with semantic concepts through the use of tagged images is a common task in social media mining. This process involves training models to learn the relationship between the visual characteristics of images and the semantic labels or keywords associated with them. Merler et al. [11] proposed a method for extracting user attributes, particularly gender information, from social media feeds. By considering the distribution of semantics in the entire set of pictures posted by a user, they aim to provide a more comprehensive and nuanced understanding compared to traditional methods that focus on text analysis or single-profile pictures. The authors in [12] addressing fundamental challenges in machine learning related to understanding and bridging the gap between semantic and visual representations, especially in the context of large-scale concept space learning. The proposed solution involves utilizing a higher-level semantic space with lower dimensionality. This is achieved by clustering correlated keywords into topics within the local neighborhood. Moreover, the authors in [13] focused on improving image annotation through a semi-automatic approach. The use of a label transfer mechanism and a specific method for image representation (sparse coding-based spatial pyramid matching) are highlighted as key components of the proposed technique. The claim of outperforming existing methods is backed by experimental results on two benchmarks, indicating the potential significance of the proposed approach in the field of image annotation.

3 Approach and Problem Definition

3.1 Semantic-Based Approach

Social user mining research seeks a deeper understanding of the semantic content within user-contributed multimedia data. However, manual image annotation is time-consuming and challenging for users to provide comprehensive tags for each image. Consequently, a semi-automatic image tagging system has emerged as a solution. In our efforts to enhance tag quality, we employed the Akiwi system, a semi-automatic image tagging tool that suggests keywords for images. The primary objective of this semi-automatic tagging system is to assign relevant keywords to images, thereby reflecting their semantic content and ultimately improving tag quality through the utilization of image content. In addressing gender classification, we adopt a semantic-based approach that leverages the keywords collected from the Akiwi system. Akiwi employs visual search algorithms sourced from Pixolution to suggest keywords, drawing upon a dataset of 22 million images from Fotolia. Figure 2 shows an example of caption in our approach, which is generated using semi-automatic tagging system techniques.

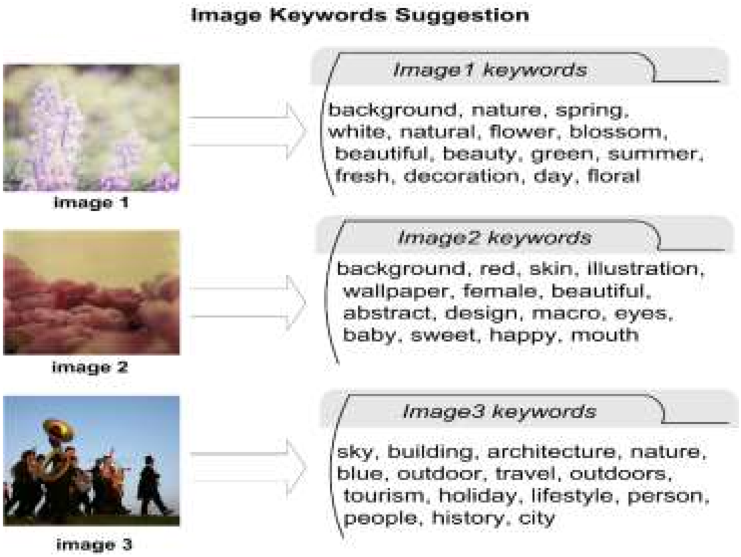
Image Keywords Suggestion
image 1
Image1 keywords
background, nature, spring, white, natural, flower, blossom, beautiful, beauty, green, summer, fresh, decoration, day, floral
image 2
Image2 keywords
background, red, skin, illustration, wallpaper, female, beautiful, abstract, design, macro, eyes, baby, sweet, happy, mouth
image 3
Image3 keywords
sky, building, architecture, nature, blue, outdoor, travel, outdoors, tourism, holiday, lifestyle, person, people, history, city

Fig. 2. Example of semantic caption in our approach

3.2 Gender Identification Definition

The problem of social user mining, as introduced in the context of Flickr and gender identification, involves the analysis of user-generated content, specifically tags and images, to determine the gender of users. Unlike a previous approach that relies solely on tags for gender identification as in [14], our proposed module incorporates both tags and images to enhance the accuracy of gender classification based on semantic contents. The problem of social user mining can be defined as follows:

Problem Definition. For a user u, given his $\mathrm{X_u}$ (multimedia objects) from Flickr, we predict the gender of u based on his multimedia objects.

For the multimedia objects, we extract features mainly represented by images or videos. To represent the multimedia feature, we label the semantic content of the user's images with a set of keywords using a semi-automatic image tagging system. This feature is denoted as:

$$X_u = (x_1, x_2, \dots, x_n), \tag{1}$$

Where X_u the user data, such as tags and images.

We hypothesize that there are distinct vocabularies associated with keywords for males and females, and this divergence can serve as a means to discern gender. To validate our hypothesis, we constructed a dictionary containing keywords vocabulary specific to both females and males. Assessing the significance of a keyword within a gender vocabulary involves tallying the occurrences of that keyword by users of the corresponding gender. We then determine the probability of a gender based on the keywords used. Table 1 presents an example of the gender dictionary tested using a sample Flickr dataset.

Table 1. Keywords Dictionary

Keyword	Male Frequency	P(male/keyword)	Female Frequency	P(female/keyword)
Soft	2012	0.452	2436	0.548
Police	4350	0.728	1623	0.272
Sisters	1108	0.408	1609	0.592
Panorama	6921	0.785	1896	0.215
Cupcakes	776	0.309	1738	0.609
Lake	9887	0.628	5869	0.372
Fisherman	2125	0.67	1045	0.33
Piazza	1085	0.679	514	0.321
Dessert	1815	0.442	2290	0.558

4 Experimental

4.1 Data Set

Flickr is an online photo and video hosting platform that allows users to share and manage their media content. Flickr offers a comprehensive API that allows developers to interact with the platform programmatically, enabling them to retrieve information, upload photos, and perform various other actions. To assess the effectiveness of the proposed algorithm, we constructed a reference dataset comprising 215,000 users. Our approach involved retrieving profile details from Flickr users using a crawler. Among these, we successfully gathered gender information for 149,700 users. The Flickr public API facilitated the acquisition of both textual and visual data, allowing us to download information with user authorization. Consequently, we obtained tags and images for the identified 149,700 users. Table 2 shows more details about our data set.

Table 2. Details of the data set

Data type	Quantity
Reference dataset	215,000 users
Ground truth	149,700 users with known gender
User's tags	Up to 350 tag per user
User's images	Up to 60 images per user

4.2 Classification Algorithm

In the context of advancing social user mining, a range of mining techniques is available. To tackle the gender classification challenge, we have chosen two widely utilized classifiers: Naive Bayes and Support Vector Machine (SVM).

The Naive Bayes classifier stands out as an exceptionally efficient and effective inductive learning algorithm in the realms of machine learning and data mining [15]. These methods are a collection of supervised learning algorithms grounded in the application of Bayes' theorem, incorporating the "naive" assumption of independence among all feature pairs. In this experiment, we employ Python tools from the Scikit-Learn library [16]. Two distinct classification methods, namely Naive Bayes and Support Vector Machine (SVM), were applied. Specifically, we employed the multinomial Naive Bayes model in this investigation, implementing the Naive Bayes algorithm tailored for multinomial distributed data, where the data is conventionally represented as a feature vector. As for SVM, we utilized the SVC (Support Vector Classification) method, implemented based on libsvm. For both classifiers, the fit(X, Y) method was employed to train the classifier with the provided training data. Subsequently, the predict(X) method was used to classify a sample of X. In this context, X denotes the feature matrix of the data, and Y represents the user label.

Given the gender classification problem having G classes$\{g_1, g_2\}$with probabilities $P(g_1)$ and$P(g_2)$, we assign the class label G to a Flickr user u based on the feature vector $X_u = (x_1, x_2, \dots, x_n)$, where x_n represent the user data, such as tags and images:

In addressing the gender classification problem with G classes and respective probabilities P(G), we assign the class label G to a Flickr user u based on the feature vector X_u, which comprises user data such as tags and images:

$$G = \arg\max_{g} P(G|X_u) \tag{2}$$

Equation (2) aims to assign the class with the maximum probability given the user data feature vector X_u. This probability is derived using Bayes' theorem:

$$P(G|X_u) = \frac{P(G) \times \prod_{i=1}^{N} P(X_i|G)}{P(X_u)} \tag{3}$$

The goal is to predict the most probable class for user *u* based on the feature vectorX_u contains N features.

Support Vector Machine (SVM) is a widely-used machine learning method for classification and various learning tasks [17]. In our experiment, we adopted the C-Support Vector Classification (SVC), which is implemented based on libsvm [18]. The core concept of applying SVM in classification is to identify a maximum-margin hyperplane that effectively separates classes in the feature vector space. Given a set of relevant Flickr data X_u, that is relevant to a user u and class labels for training$\{(X_u, G)|u = 1, \dots, n\}$, where X_u represent the feature vectors of user data and G is the target class label, the SVM will map these feature vectors into a high dimensional space.

4.3 Experimental Result and Discussion

To assess the effectiveness of our method, we evaluate its performance using metrics such as classification accuracy (Acc), precision (Pre), recall (Rec), and F1 score, as defined by the following equations:

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{4}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{5}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{6}$$

$$\mathrm{F1} = 2\left(\frac{\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre+Rec}}\right) \tag{7}$$

Here, TP represents true positives, TN stands for true negatives, FP denotes false positives, and FN indicates false negatives.

Table 3. Experiment result for semantic-based approach.

Features	Approach	Acc	Pre	Rec	F1
Keywords	NB	0.84	0.83	0.85	0.83
	SVM	0.84	0.85	0.84	0.82
Tags	NB	0.80	0.84	0.80	0.80
	SVM	0.76	0.57	0.76	0.65
Keywords+Tags	NB	0.82	0.82	0.82	0.81
	SVM	0.80	0.64	0.80	0.70

The experimentation involved sampling the dataset across various features and classifiers, and then evaluating each classifier and feature's performance. The results are presented in Table 3 above. Notably, the gender classification achieved an accuracy exceeding 84% when utilizing keywords with both classifiers, indicating the superiority of the proposed semantic-based approach over the content-based alternative. In terms of classifiers, Naive Bayes demonstrated a slight superiority over SVM, with this preference attributed to its ability to perform well even in the presence of some missing data. The results of evaluating our approach's performance using metrics such as classification accuracy (Acc), precision (Pre), recall (Rec), and F1 score are illustrated in Figure 3.

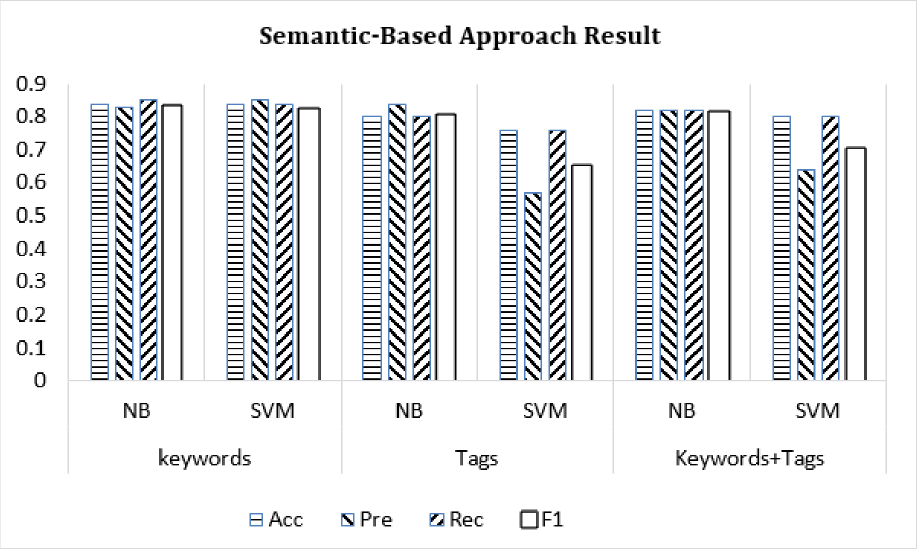

Fig. 3. The evaluation result of the performance using metrics such as classification accuracy (Acc), precision (Pre), recall (Rec), and F1 score.

5 Conclusion

In conclusion, this paper successfully addresses the challenge of identifying user gender through a semantic-based data technique and the implementation of a semi-automatic image tagging system. The utilization of Naive Bayes and Support Vector Machines (SVM) as classification algorithms, with data represented as feature vectors, yields promising results. The experimental findings, based on a substantial sample of over 149,700 Flickr users, demonstrate an impressive accuracy of over 84% in gender identification. This underscores the effectiveness of the combined Naive Bayes and SVM algorithms, emphasizing their potential for accurate gender classification in user-generated content scenarios. The presented approach contributes valuable insights to the field of gender identification, showcasing a practical and efficient methodology for addressing this challenge.

References

1. C. Eke, A. Norman, L. Shuib, and H. Nweke, “A survey of user profiling: State-of-the-art, challenges, and solutions,” *IEEE Access*, vol. 7, pp. 144907–144924, 2019.
2. C. Peersman, W. Daelemans, and L. Van Vaerenbergh, “Predicting age and gender in online social networks,” in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, ser. SMUC ’11. New York, NY, USA: ACM, 2011, pp. 37–44.
3. J. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on twitter,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser.

EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1301–1309.

4. K. Filippova, “User demographics and language in an implicit social network,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL ’12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 1478–1488.
5. M. Cheung and J. She, “An analytic system for user gender identification through user shared images,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3, pp. 30:1–30:20, Jun. 2017.
6. R. Miura, M. Hirota, D. Kato, T. Araki, M. Endo, and H. Ishikawa, “Predicting user gender on social media sites using geographical information,” 2018.
7. J. Shafi, R. Adeel Nawab, and P. Rayson, “Semantic tagging for the urdu language: Annotated corpus and multi-target classification methods,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 6, Jun. 2023.
8. S. Xiang, D. Qian, J. Gao, Z. Zhang, T. Liu, and Y. Fu, “Rethinking person re-identification via semantic-based pretraining,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–17, 2023.
9. L. Wu, S. Hoi, R. Jin, J. Zhu, and N. Yu, “Distance metric learning from uncertain side information for automated photo tagging,” vol. 2, pp. 1–28, 2011.
10. P. Wu, S. Hoi, P. Zhao, and Y. He, “Mining social images with distance metric learning for automated image tagging,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM ’11. New York, NY, USA: ACM, 2011, pp. 197–206.
11. M. Merler, L. Cao, and J. Smith, “You are what you tweet… pic! gender prediction based on semantic analysis of social media images,” in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.
12. M. Wang, X. Zhou, and T.-S. Chua, “Automatic image annotation via local multi-label classification,” in *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, ser. CIVR ’08. New York, NY, USA: ACM, 2008, pp. 17–26.
13. W. Zhang, Z. Qin, and T. Wan, “Semi-automatic image annotation using sparse coding,” in *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*, vol. 2, July 2012, pp. 720–724.
14. A. Popescu, G. Grefenstette *et al.*, “Mining user home location and gender from flickr tags.” in *ICWSM*, 2010.
15. H. Zhang, “The optimality of naive bayes,” *A A*, vol. 1, no. 2, p. 3, 2004.
16. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
17. N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
18. C. Chang and C. Lin, “Libsvm: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

تحديد الجنس على أساس الدلالات من خلال المحتويات التي ينشئها المستخدمون

محمد علي إبراهيم الطاهر[1]

1 جامعة طرابلس كلية تقنية المعلومات
m.eltaher@uot.edu.ly

الملخص: إن جنس المستخدم هو معلومات بالغة الأهمية للخدمات والتطبيقات المخصصة في الشبكات الاجتماعية عبر الإنترنت. وهو يؤثر على مجالات مثل أنظمة التوصية والإعلان واكتشاف الاتصال. ومع ذلك، قد تكون معلومات جنس المستخدم مخفية أو غير محددة في الشبكات الاجتماعية عبر الإنترنت، مما يؤدي إلى عدم الدقة أو القيود في تطبيقات مختلفة. تساهم التفاعلات اليومية لمليارات المستخدمين على الشبكات الاجتماعية عبر الإنترنت مثل Flickr في إنشاء كميات هائلة من المحتوى الذي ينشئه المستخدم. يتضمن هذا المحتوى وسائط متعددة مثل الصور ومقاطع الفيديو والمعلومات النصية. الهدف الأساسي من هذه الورقة البحثية هو معالجة تحدي تحديد جنس المستخدمين متضمناً نهجن تقنية بيانات قائمة على الدلالات. إن استخدام نظام وسم الصور شبه التلقائي يعني عملية يتم فيها تصنيف الصور باستخدام الأتمتة، مما قد يحسن الكفاءة والدقة. نستخدم خوارزميتين للتصنيف لتحديد الجنس: Naive Bayes وSVM ، حيث يتم تمثيل البيانات عادةً كمتجهات مميزة. تظهر نتائجنا التجريبية على أكثر من 149700 مستخدم لـ Flickr دقة تزيد عن 84٪ لتحديد الجنس. ويشير هذا إلى أن الجمع بين خوارزميات Naive Bayes وSVM ، مع تمثيل البيانات كمتجهات مميزة، أثبت فعاليته في تصنيف الجنس بناءً على المحتوى الذي ينشئه المستخدم.

الكلمات المفتاحية: تحديد الجنس ، تحديد ملف تعريف المستخدم ، المحتوى الدلالي ، المحتوى الذي ينشئه المستخدم.