|  |  |
| --- | --- |
| **Title** | Twitter gender classification using user unstructured information |
| **Date** | 2015 |
| **Authors** | Marco Vicente  Fernando Batista  Joao Paulo Carvalho |
| **Summary** | 1)This project proposes a method to automatically detect the user’s gender (male or female), based on unstructured information extracted from the user’s profile, and made available by Twitter for each tweet.  2)The only restriction for this method is that within the user profile there is at least a sequence of characters matching a name contained within a dictionary.  3) A set of manually defined features are proposed for extracting useful information from the user’s profile attributes, namely user name, and screen name.    4) the online content generated by the user in each tweet is not used in the scope of this work. Attributes, such as the user name, commonly encode relevant information about the gender of the user.  5) We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that colors alone can provide reasonably  6) that may provide additional information about the user in an unstructured form: • Screen name (e.g.: johndoe95) • User name (e.g.: John Doe the best :) ) • Location • URL • Description  7) We have described an approach to automatically detect the gender of Twitter users, based on clues provided by their profile information. Several name-related features that capture phenomena specific to Twitter users are proposed and evaluated on a dataset of about 242K English users. Different supervised and unsupervised approaches are used to assess the performance of the proposed features, including Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and K-means |
| **Accuracy** | CLASSIFICATION RESULTS FOR SUPERVISED METHODS:   1. Logistic Regression🡪 94.5% 2. Multinomial Naive Bayes🡪 97.2% 3. Support Vector Machines🡪 96.0%   CLASSIFICATION ACCURACY FOR UNSUPERVISED METHODS:   1. K-means clustering🡪 74.9% 2. Fuzzy c-Means🡪 84.9%   **Overall ACC**🡪 96% |
| **Title** | Language-Independent Gender Classification on Twitter |
| **Date** | May 2013 |
| **Authors** | Jalal S. Alowibdi1, 2, Ugo A. Buy1 and Philip Yu1, 2 |
| **Summary** | 1. Gender classifications typically are language-dependent, not scalable, inefficient, held offline using high-dimensional spaces. A recent study [1] shows that there are around 78 different languages on Twitter with English as the dominant language. Another study by Wauters shows that only around 50% of Twitter messages are in English1. Our Twitter dataset alone contains 31 different languages. 2. We defined a novel approach for predicting gender using color-based features. Our method is language independent; most other existing methods that use text are restricted to one language or a few languages. 3. We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that colors alone can provide reasonably accurate gender predictions. 4. We defined a color quantization and sorting technique for preprocessing colors harvested from Twitter profiles. This technique substantially improves prediction accuracy. 5. A Twitter user must first fill a profile form, consisting of about 30 fields containing biographical and other information, such as personal interests and hobbies. However, many fields in the form are optional, and indeed substantial portions of Twitter users leave many or all of those optional fields blank. In addition, Twitter’s profile form does not include a specific “gender” field, which complicates gender identification for Twitter users. |
| **Accuracy** | ACCURACY OF GENDER PREDICTIONS FOR DATASET T3 WITH RGB COLORS WITHOUT QUANTIZATION.:  1)NB🡪 61.2%  2)DT🡪63.3%  3)PNN🡪59.4%  4)NB-Tree🡪65.7%  ACCURACY OF THE EXPERIMENT RESULTS FOR DATASET T3 AFTER APPLYING COLOR QUANTIZATION AND SORTING:  1)NB🡪 56.0%  2)DT🡪64.7%  3)PNN🡪69.0 %  4)NB-Tree🡪71.4%  **Overall ACC**🡪 66% |
| **Title** | Gender Classification using Twitter Text Data |
| **Date** | June 2020 |
| **Authors** | Pradeep Vashisth, Kevin Meehan |
| **Summary** | 1. Gender classification can be treated as a two-class or binary classification problem in practice for example given two classes, male or female. This classification aims to assign the anonymous text or message to one of the given classes without knowing about the user. 2. The issue with gender detection using Twitter data has not previously been searched 3. at scale. Although similar NLP research has been conducted to automatically detect the gender of a writer of the text, Advancement in ML and text classification techniques has helped the growth of research within this social media domain. 4. The research focused on detecting gender, age, political orientation, and nationality. Using status updates, communication behavior, and social network structure, the author aimed to research the usefulness of language content processing. The aim was to discover latent user attributes to automatically detect some of the latest features in the text. 5. Deitrick researched detecting gender classification by using neural network techniques. This research used a manual label dataset of 3,031 tweets. Both, balanced winnow and modified winnow techniques were evaluated. The research achieved 98.5% accuracy by applying both models and 53 n-gram features selection 6. Naïve Bayes, Perceptron, and word anagram were applied during this experiment. This model achieved significant results and achieved 99.3% accuracy by using the Perceptron while tweets length was at least 75 characters. |
| **Accuracy** | Baseline (TF-IDF) approach  1)LR (Logistic Regression) 🡪53.65 MLP  2)(Multilayer Perceptron) 🡪48.75  3)SVM (Support Vector Machine) 🡪52.67  4)Naïve Bayes🡪 53.84  5)Random Forest🡪 47.7  6)XG-Boost 🡪54.93  Word Embedding (W2Vec) approach:  1)LR (Logistic Regression) 🡪57.14  2)SVM (Support Vector Machine) 🡪52.67  3)Random Forest 🡪47.72  4)XG-Boost🡪 55.38 |