ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Identification of Author of Text Using Stylistic Analysis with the Help of Twitter Data

Aishwarya Bhonde, Hasna Mohammed, Rohit Keshwani, Prachi Sarode

Student, Dept. of I.T, MIT College of Engineering, Pune, India

Student, Dept. of I.T, MIT College of Engineering, Pune, India

Student, Dept. of I.T, MIT College of Engineering, Pune, India

Assistant Professor, Dept. of I.T, MIT College of Engineering, Pune, India

ABSTRACT: Stylistic Analysis is a discipline that determines authorship of literary works through the use of statistical analysis and machine learning. While this discipline has been used successfully to determine authorship of famous literary works, the area of analyzing digital content is still relatively new with much more to discover. It is a behavioral feature that a person exhibits during writing and can be extracted and used potentially to check the identity of the author of online documents. Since the early to mid-1990's the explosion of the Internet has opened up new uses for stylometry in the area of email, social networking and blogging websites etc. This project is developed in three parts. Part I of the project focuses on data extraction from Twitter. Part II of the project is Data Pre-processing and Feature Extraction from the acquired data in Part I. Part III includes classification based on the features extracted in Part II using classification algorithms-Naïve Bayes. Once the classification is done, the best few possible matches of the profiles of people are displayed as the final result of the system.

KEYWORDS: Stylometry, Naïve-Bayes, Extraction, Pre-processing, tweepy, classification.

I. INTRODUCTION

Stylometry is the behavioural feature that a person exhibits during writing and can be extracted and used potentially to check the identity of the author of online documents. When someone authors a literary work, document, or email they leave behind certain attributes to their writing style that can be analysed and used to determine other works by the same author. Some of these attributes or features are vocabulary usage, sentence complexity, specific phrases, and many others [1], [2].

Trying to determine authorship of digital or Internet content presents some different and unique challenges that were not introduced with conventional stylometry. One of the key challenges is number of authors. Many of the previous studies such as authorship identification of Shakespeare's works and the Federalist Papers, dealt with a relatively small number of potential authors, typically no more than 10.

Digital content such as electronic mail and social networking applications like Twitter could have more than 10 authors and sometimes hundreds of potential authors. For example, if someone is planning a crime utilizing Twitter they most likely would not use their real name or any accurate characteristics to identify themselves. After narrowing their search for an original author using other methods this could still leave law enforcement with a large pool of potential authors.

Another significant challenge working with digital content is length of the work. For example, email and blogging content tend be a lot shorter than some of the previously mentioned works. Trying to attribute authorship to content of less than 250 words is more difficult and challenging [14], [15].

Analyzing all these flaws of the conventional stylometry, the proposed stylometry tries to eliminate all these with the help of Natural Language Processing (NLP) used coupled with the machine learning. To perform such stylometry, there would be a requirement of huge data (live data) i.e. big data that would be obtained briefly from websites such as twitter. The live data can be easily obtained from these developer websites with the help of API's provided by default by these websites [3].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

The obtained data would be stored in the data warehouse as the data would be huge. The processing of this is done and once the match is found, the required data is then feature extracted to the backend database. The feature extracted data is then classified using classifiers such as Naïve Bayes classifier. As a result, after the classification is done, the more precisely accurate classified output will be preserved and others are dropped.

II. LITERATURE SURVEY

i. Analyzing the social media:

With the course of time, there has been a huge increase in the number of people using internet. With the growth of internet, access to social media has also grown. So the social websites such as Facebook, Twitter, Wordpress, etc has led more people to connect to them and share their views [11], [12]. With such an easy access to the internet, multiple accounts by a user on social media have also increased letting to its misuse. This questions the authenticity of authorship of a particular post, tweet, blog, etc. So, the verification of authenticity of authorship of these posts, tweets, blogs, etc has become important.

ii. System configuration:

After mining the social media websites all the data extracted requires a storage facility, considering the requirement of an efficient database, a NoSQL database fulfils it [5]. For the extraction of such data, the latest version of python i.e. Python3 has a simple approach to coding, considering the simplistic connectivity between python and a NoSQL database i.e. MongoDB, the system requirement is configured.

iii. Classifier:

The verification of authorship of the text is done through its classification on the basis of particular inputs. For this, a classifier that can handle multiple classes i.e. inputs is required. Naïve Bayes classifier is a multi-class classifier that is preferable for the classification.

III. TECHNOLOGIES USED

i. Python3:

In this project, Python3 is used for different processing of data. Python is used for Natural Language processing and classification, as it is easier through python. Data pre-processing such as tokenizing and removal of stopwords is done through nltk kit available in python. Python also provides interfaces to all major commercial databases.

ii. MongoDB:

MongoDB is used as database to store the extracted tweets and profile information from twitter. Tweets and screenname are accessed from this database to train the classifier. MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

iii. Tweepy:

The Twitter APIs provide programmatic access to read and write Twitter data. Tweepy is the free API provided by twitter for accessing the public profiles information from twitter. An easy to use python library is available for accessing the twitter API. For accessing through API required credentials are provide for authentication.

iv. Flask:

Flask is used as web framework for connectivity with python and HTML/CSS. It provides tools, libraries and technologies that allow building web application.

IV. SYSTEM ANALYSIS

i. Online Data Extraction:

To get the datasets for processing the data has to be extracted from web. Various social networking websites such as Twitter, Facebook, and Google+ are available for data extraction. The data from Facebook that can be extracted is in the form of statuses, posts, comments, etc. Twitter boasts of various tweets, retweets and shares from millions of users [18]. Gmail allows developers to access certain mails for analytical purposes that are under their rules and regulations, with the help of Google developer API, numbers of mails that are allowed for extraction can be extracted [9].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

For our application we are extracting tweets from twitter, using freely available API provided by Twitter-Tweepy. All the data extracted from these sources are stored in the data warehouse. The data warehouse will be constructed with the use of MongoDB. The data warehouse will consist of the extracted data from where the data will be taken for further processing.

ii. Data Pre-Processing:

Data pre-processing is one of the required step of Stylistic Analysis. Data extracted from online are not in the suitable form for processing and train the classifier. It should be converted to suitable format for quality results. The data pre-processing of online data involves tokenization, removal of URLS or links and Removal of Stopwords.

a. Tokenization:

Tokenization is the process of splitting the input sentence into meaningful units, these units are called tokens.
A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing [13].

b. Removal of URLs:

Online data extracted may have various links or URLs Included in the tweets or posts. Before training the classifier all such redundant elements should be removed from the text [7]. The tokenized elements should be checked if it has any links and it should be removed to get link free text.

c. Removal of Stopwords:

The removal of Stopwords or high frequency tokens after tokenization is one of the important steps in data pre processing. In this step each tokens are compared with predefined high frequency stop word defined in NLTK and removed from input text [9]. The advantage of this step is that, it reduces the text size which helps to reduce time complexity and space complexity in training the classifier.

iii. Feature Extraction:

The data that is acquired into the data warehouse from various online sources is now ready for feature extraction. A particular user can be classified on the basis of humongous features. But there are certain features out of these that will uniquely identify the authorship of the text. These features that are required to classify a profile are stop words, punctuations, vocabulary richness, unique words, etc. On the basis of these feature extraction, the training set is created.

The Natural Language Processing (NLP) toolkit of python provides various functionalities using which the particular module of a feature can be created [5]. Using this toolkit, all the mentioned features can be given an actual implementation. Every individual feature module is then combined to form a training set.

iv. Training Set:

The training set is generally made manually, so that we can train the model in the way we need it. Creating a training set can be done in multiple ways like we can create it on our own or even crowd source it. Based on the features extracted, a training set is built for the algorithm for each person's profile. The training set is created by combining all the features that are extracted in the previous module.

The training set is so called because the developers can train the set to behave as per the requirement of the system. The training set consists of the extracted data that will finally be provided to the classifiers to distinguish among the people. All the data that is extracted from all the websites that resides in the data warehouse is now distinguished into the training set.
This distinguished data will be stored in the primary backend database created using MongoDB. This data will be provided to the classifier that will be essential to classify the test set as per the behaviour of the training set which ultimately gives the profile of the person.

v. Classification:

The classifier module has two inputs – training set and the test set. The training set is the developed set from the various features described in the previous step. The test set is the input from the registered user that tries to identify the authorship of the text input. The classification of the test set is done on the basis of the training set in order to generate

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

the required output. The classifiers used for classification is Naïve Bayes Classifier. Once the classification is done, the identification of the required person is confirmed and top n number of matches is given as the output.

a. Naïve Bayes Classifier:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [7].
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
Bayes theorem is used to calculate posterior probability P(c|x)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Where,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the posterior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor [1].

Naïve Bayes Classifier is used because it is easy and fast to predict class of test data set as well as also perform well in multi class prediction which is needed for this application. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

vi. Result:

Once the classification is done the result is given as class to which the input text belongs, here every user whose tweets has been extracted is considered as a class. It also gives the probability of match with input text to each class. The returned result can be used to identify the accuracy level of the classification.

V. SIMULATION RESULT

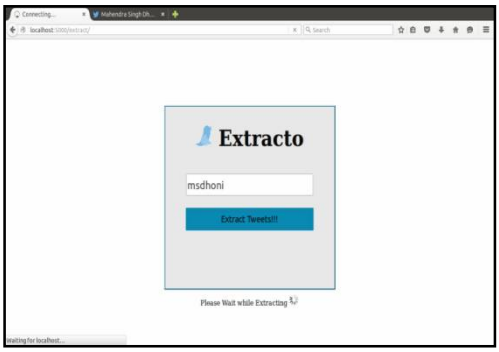

Fig 1: Twitter Extraction Model – 1

The simulation of the stylistic analysis system starts with the data extraction from twitter as shown in Fig 1. The username of the twitter account of a particular user has to be entered so as to extract the tweets.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

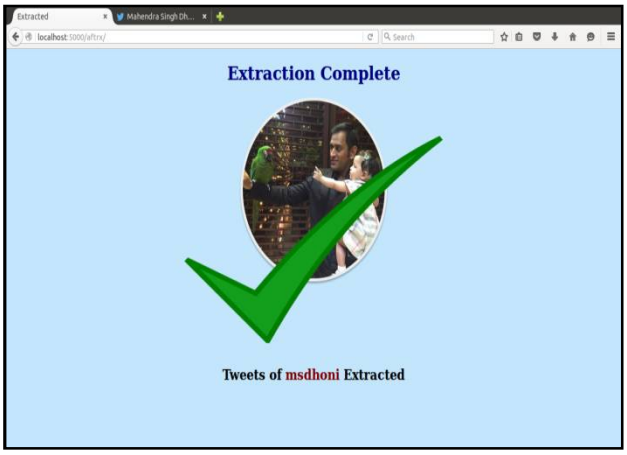

Fig 2: Twitter Extraction Model – 2

Once the extraction of the tweets are complete, the acknowledgement of completion is given as shown in Fig 2. The tweets are extracted and stored in the mongoDB data warehouse. The contents of every particular extracted tweet are publish date, profile icon, text of the tweet etc.

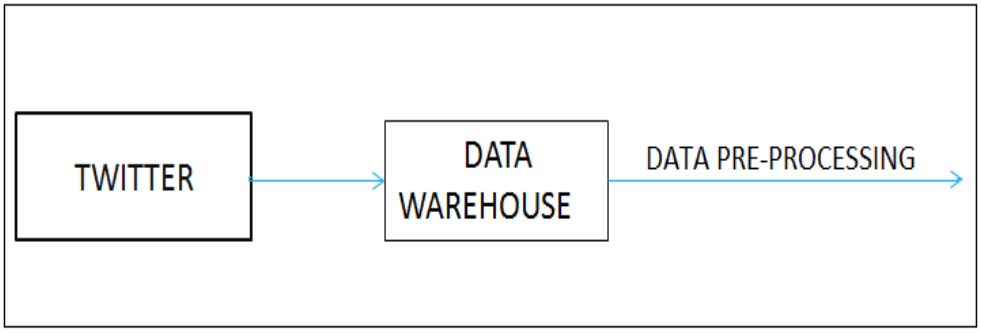

Fig 3: Online Data Extraction

Fig 3 shows the complete online data extraction model. In this, the data from Twitter is extracted and stored into the data warehouse constructed using MongoDB. This data is further sent for pre-processing.

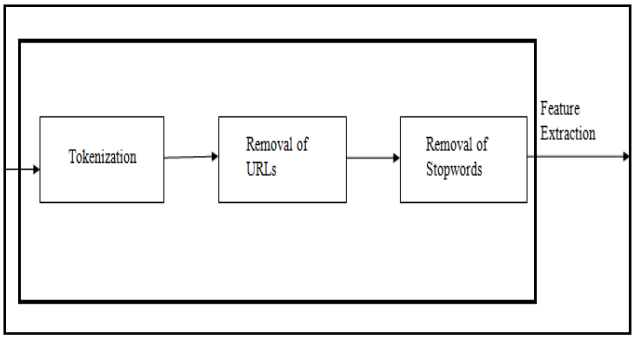

Fig 4: Data Pre-processing

The data pre-processing of the extraacted tweets takes places in the order as described in the fig 4. The tokenization of data i.e splitting of sentences into words takes palce. Then every URLs are removed as they may cause insignificant results. The stopwords or high frequency words such as i, am, the, this, that etc are also removed as they are irrelevant

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

in the identification of the author. The now pre-processed text is now sent for feature extraction where the text is split into 1-gram. Now, the training set is ready.

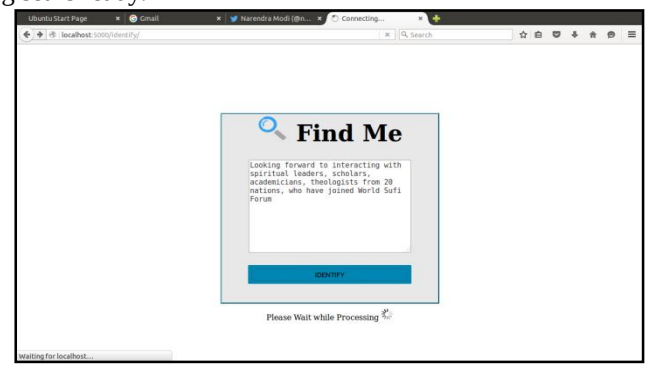

Fig 5: Data classification model -1

Once the training set is ready, the test set has to be recorded from the system GUI. This is done as shown in fig 5. The test set as some random text is accepted and the task is left upon the classifier to classify the test data on the training data genetrated before.

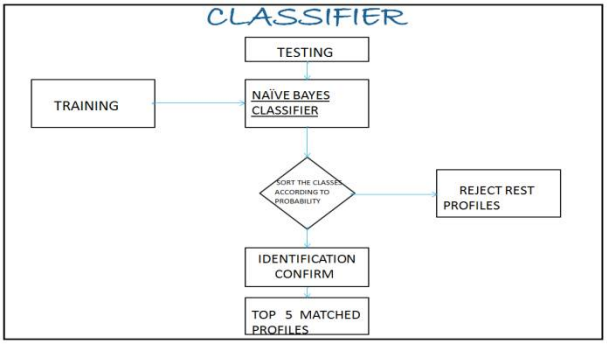

Fig 6: Classification flow

Figure 6 shows the working of the naive bayes classifier. The naive bayes cassifier accepts two inputs – i) training set and ii) test set. It then classifies the test set on the basis of training set and displays the top 5 matches on the base of highest match probability. The rest matches are rejected by the classifier.

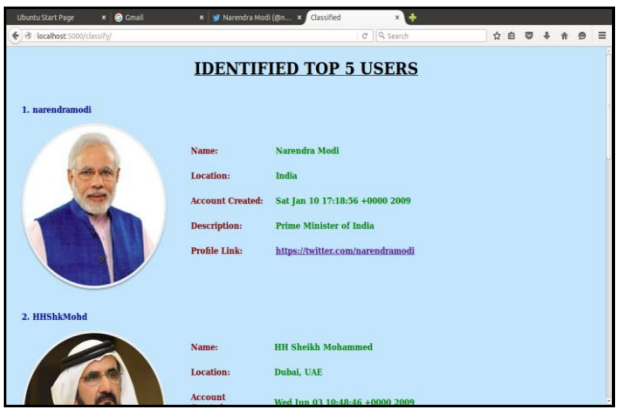

Fig 7: Data classification model – 2

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Once the classification is complete, the fig 7 shows the top 5 matches. In this way, the simulation of the system concludes with the identification of the author in the form of top 5 probable matches.

VI. CONCLUSION

In this era of internet, the use of online blogs, forum, social network and email is very popular for communication. At the same time due to anonymity, cybercriminals make use of these online messages for illegal activities like cyber bulling, fishing, etc. In this concept this application helps to find the author of the anonymous text by extracting the existing data of the users and processing that data to identify the most nearest match of the text.
The new proposed system titled, "Stylistic Analysis" is entitled to verify the authorship of the text on the basis of the enormous data collected from Twitter. The feature extraction of this data is done and the training set is created which identifies the user input in the form of test set. The data extraction is done with the help of major API provided by the Twitter such as Tweepy.
The feature extraction is done and the classification is done using Naïve Bayes classifier and the output is given in the form of the profile of the text. The more accurate top 5 classification is taken into consideration and the others are rejected.

REFERENCES

[1] Jenny S. Li, John V. Monaco, Li-Chiou Chen, Charles C. Tappert, "Authorship Authentication Using Short Messages from Social Networking Sites", 2014.
[2] Marcelo Luiz Brocardo, IssaTraore,Sherif Saad,Isaac Woungang , "Authorship Verification for Short Messages using Stylometry ,"2013.
[3] Gregory Shalhoub, Robin Simon, Ramesh Iyer, Jayendra Tailor, Dr. Sandra Westcott "Stylometry System – Use Cases and Feasibility Study" on May 7th, 2010.
[4] Ahmed Abbasi, Hsinchun Chen,"A Stylometric Approach toIdentity-Level Identification and Similarity Detection in Cyberspace," 2008.
[5] Ethem Alpaydin, Introduction to Machine Learning, PHI 2nd Edition-2013.
[6] Peter Flach, Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, Edition 2012.
[7] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst., 26:7:1–7:29, April 2008.
[8] O. Canales, V. Monaco, T. Murphy, E. Zych, J. Stewart, C. T. A. Castro, O. Sotoye, L. Torres, and G. Truley. A stylometry system for authenticating students taking online tests. CSIS, Pace University, May 6 2011.
[9] C. E. Chaski. Who's at the keyboard: Authorship attribution in digital evidence investigations. International Journal of Digital Evidence, 4(1), Spring 2005.
[10] X. Chen, P. Hao, R. Chandramouli, and K. P. Subbalakshmi. Authorship similarity detection from email messages. In Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition, MLDM'11, pages 375–386, Berlin, Heidelberg, 2011. Springer-Verlag.
[11] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. Digital Investigation, 8(1):78 – 88, 2011.
[12] N. Cheng, X. Chen, R. Chandramouli, and K. Subbalakshmi. Gender identification from e-mails. In Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on, pages 154 –158, 30 2009-april 2 2009.
[13] J. H. Clark and C. J. Hannon. A classifier system for author recognition using synonym-based features. In Proceedings of the 6th Mexican international conference on Advances in artificial intelligence, MICAI'07,pages 839–849, Berlin, Heidelberg, 2007. Springer-Verlag.
[14] M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In Proceedings of the 18th Annual Computer Security Applications Conference, pages 282 – 289, 2002.
[15] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem. Towards an integrated e-mail forensic analysis framework. Digital Investigation, 5(3-4):124 – 137, 2009.
[16] H. V. Halteren. Author verification by linguistic profiling: An exploration of the parameter space. ACM Trans. Speech Lang. Process., 4:1:1–1:17, February 2007.
[17] N. Homem and J. Carvalho. Authorship identification and author fuzzy fingerprints. In Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American, pages 1 –6, march 2011.