Prediction of Personality Traits in Twitter Users with Latent Features

Conference Paper · March 2019

DOI: 10.1109/CONIELECOMP.2019.8673242

4 authors:

Dora-Luz Almanza-Ojeda
Universidad de Guanajuato
22 PUBLICATIONS **133** CITATIONS
SEE PROFILE

Juan Carlos Gomez
Universidad de Guanajuato
49 PUBLICATIONS **542** CITATIONS
SEE PROFILE

Mario Alberto Ibarra-Manzano
Universidad de Guanajuato
89 PUBLICATIONS **660** CITATIONS
SEE PROFILE

Daniel Ricardo Jaimes Moreno
Universidad de Guanajuato
1 PUBLICATION **9** CITATIONS
SEE PROFILE

Prediction of Personality Traits in Twitter Users with Latent Features

Daniel Ricardo Jaimes Moreno
Departamento de Ingeniería Electrónica
División de Ingenierías Campus Irapuato-Salamanca
Universidad de Guanajuato, Salamanca, Mexico
dr.jaimesmoreno@ugto.mx

Juan Carlos Gomez*
Departamento de Ingeniería Electrónica
División de Ingenierías Campus Irapuato-Salamanca
Universidad de Guanajuato, Salamanca, Mexico
jc.gomez@ugto.mx

Dora-Luz Almanza-Ojeda
Departamento de Ingeniería Electrónica
División de Ingenierías Campus Irapuato-Salamanca
Universidad de Guanajuato, Salamanca, Mexico
dora.almanza@ugto.mx

Mario-Alberto Ibarra-Manzano
Departamento de Ingeniería Electrónica
División de Ingenierías Campus Irapuato-Salamanca
Universidad de Guanajuato, Salamanca, Mexico
ibarram@ugto.mx

***Abstract*—The globalization of Economy has forced the society to maintain a constant evolution in marketing techniques. It is thus very important to design tools and methods that allow knowing and characterize individuals in groups to develop effective marketing strategies. In this context, any company would be interested in finding the tastes and preferences of people regarding the products and services offered in the global market. One technique that could help in this, is the analysis of the personality of each individual to identify their tastes and preferences. In this way we can offer products and services that meet their needs through appropriate advertising for each type of personality. In this work, we propose the use of latent features, extracted with a diversity of dimensionality reduction methods, to infer the personality of Twitter users using textual content-based features, and we compare the performance of the different techniques. For conducting our experiments, we use the PAN CLEF 2015 dataset consisting of 14,166 tweets in English of 152 different users, and a diversity of classification methods. Our results shows interesting insight about the personality prediction task.**

***Index Terms*—Twitter, latent features, dimensionality reduction techniques, personality.**

I. Introduction

Research in psychology suggests that by means of personality traits, we can largely explain the behavior and preferences of people [1]. In turn, knowing the behavior and preferences of persons, allows us to improve recommendations systems considerably [2] to present better suited content, such as websites, products, brands and services [3].

In psychology there are several typologies of personality traits, but the best known and more used is the Big Five typology [4]. Psychologists generally apply this model manually through a questionnaire for their patients. The model evaluates and analyzes the composition of five dimensions of personality in its broadest sense. The five main traits or factors are traditionally referred to as: openness (openness to new experiences), conscientiousness (responsibility), extraversion, agreeableness (kindness) and neuroticism (emotional instability).

The questionnaires for measuring the personality traits in the Big Five model contains between 20 to 360 questions; having a large variation in content depending on the test and the focus of the researcher. Answering a large questionnaire requires a lot of time and it is tedious for users, especially in the context of online services, where most of the users may not be willing to fill out the form completely, because of the long and boring process.

Nevertheless, recently several methods have emerged to demonstrate that it is possible to automatically infer to a certain degree the personality of the users in online services. For example, Lambiotte et al. [5] and Youyou et al. [6] exposed in their research that the automatic prediction of personality based on the likes of Facebook is more accurate than predictions made by the users' friends, or even those of their partners.

On the other hand, personality prediction is one of the most difficult author profiling tasks. Author profiling (AP) consists in predicting or inferring as much knowledge as possible about an unknown author, simply by analyzing a specific text written by him/her [7], [8]. The interest of the scientific community for the tasks of AP has increased in recent years. This increase is due, in part, to the enormous amount of textual information generated by users on the internet [9] and specially in social media, such as Facebook and Twitter.

In the case of Twitter, the users in the site posts mainly text (tweets). This is relevant to consider Twitter as a social network suitable to conduct research on automatic personality detection. In our case, we used for our study the English PAN CLEF 2015 dataset, consisting of 14,166 English tweets corresponding to 252 different users.

There are several ways to approach the task of personality prediction of users. One of them is to approach it as a classification problem. Generally speaking, in text classifica-

* Corresponding author.

tion tasks, there are three important steps: 1) extraction of textual features, 2) the representation of documents, and 3) the application of a learning algorithm [10].

In this work we focus on the extraction of features, by using three different dimensionality reduction methods, Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA) and Non-negative Matrix Factorization (NMF), to extract latent features as linear combinations of superficial features, and used them to predict the personality traits from Twitter users. It is important to notice that such techniques have not been tested in the literature to predict personality traits. We conduct several experiments with each type of latent features, and with combinations of these, using a diversity of classifiers, and we compare their performances. Our results produce interesting insights about the task.

The rest of the paper is organized as follows. In Section II we present some relevant related work for personality traits prediction in social media. In Section III we present the methodology we followed to solve the task, including the description of the PAN CLEF 2015 dataset. In Section IV we show the results of our experiments. Finally, in Section V we summarize the conclusions and some ideas for future research.

II. Related work

In this section we present some published papers that support this study. In particular, we show the most modern advances related to personality prediction for Twitter users.

Pervaz et al. [11] expose that the main approaches used for the automatic identification of an author's personality traits from text can be classified into three broad categories: 1) stylometry-based approaches (which aim to identify the features of an author from their writing style), 2) content-based approaches (which identify the features of the author using features extracted from the content of the document) and 3) subject-based approaches (which attempt to predict the profile of an author based on the themes used in the document).

Grivas et al. [12] propose the split of stylometric and structural features for the task. The group of structural features represent basically counts of specific Twitter features, for example: mentions (ats), hashtags and URLs in the users' tweets. For the capture of stylometric features, they based their approach on trigrams, arguing that trigrams capture stylometric features well and are more extensible to unknown text when using a small training set, comparing to a bag of words approach. Regarding personality prediction, they treated it as a regression problem using support vector machines.

González et al. [13] present in their work a method that combines the stylistic features represented by the n-grams of characters and the n-grams of grammar labels (part-of-speech) to solve the problem of author profiling. The authors applied in both groups of n grams a context-dependent dynamic normalization to extract as much stylistic information as possible from the documents. With the n-grams of characters they extract several stylistic elements: frequency of characters, use of suffixes (gender, number, tenses, diminutives, superlatives, etc.), use of punctuation marks (frequency of use, repetition), use of emoticons, etc [14], [15]. The authors explain that the POS n-grams provide information regarding the way in which the text is structured: the frequency of grammatical elements, the diversity of grammatical structures used and the interaction between grammatical elements.

The method proposed by Poulston et al. [16] employs a machine learning approach to predict the personality of users on Twitter. He used topic models, implemented using Latent Dirichlet Allocation, and n-gram language models to extract features to train several regressor models.

Some researchers for the automatic identification of an author personality traits using text, make use of dictionaries from psychology studies. In the case of Arroju et al. [17], their work is based on the dictionary of linguistic consultation and counting of words (LIWC). In their work, they explain how they associate the words from text with those of the LIWC dictionary, and the way in which the feature vector was created to represent each user of the dataset.

Finally, Alvarez et al. [10]. propose using dimensionality reduction techniques in the upper part of the descriptive and discriminative textual features typical for the AP task. Specifically, they propose the joint use of the techniques of Second Order Attributes (SOA) and Latent Semantic Analysis (LSA) to highlight discriminative and descriptive properties, respectively. Their experimental results in AP show that the combination of SOA and LSA exceeds a bag-of-words and each individual representation, demonstrating its usefulness in predicting gender, age and personality profiles.

Beside the works presented in [10] and [16], there are not other works where dimensionality reduction techniques have been used to extract latent features for predicting personality traits. In our work we study the performance of three other methods, PCA, LDA and NMF, that have not been explored for the task.

III. Methodology

A. Dataset setup

For the experiments in this work we used the training dataset from the PAN CLEF 2015 conference[1]. The dataset contains 14,166 tweets from 152 users, with around 100 tweets per user. The tweets are in English and the content speaks about a diversity of topics, with a vocabulary of 17,369 different words. The dataset is grouped by user and labeled with the values of the 5 personality traits per user: openness, conscientiousness, extraversion, agreeableness and neuroticism. The values of each personality trait are between -0.5 and 0.5. -0.5 indicates the total absence of the trait, and 0.5 indicates a strong presence of the trait. Table I shows the average number of words in tweets from the different personality traits in the PAN CLEF 2015 dataset. In this table we observe that the extroverted tweets contain more words on average that the tweets from other traits, which is something expected.

[1] Dataset available at: https://pan.webis.de/clef15/pan15-web/author-profiling.html

TABLE I
PAN CLEF 2015 DATASET: AVERAGE NUMBER OF WORDS IN TWEETS FROM DIFFERENT PERSONALITY TRAITS

Value	Extroverted	Stable	Agreeable	Conscientious	Open
-0.5	0	0	0	0	0
-0.4	0	0	0	0	0
-0.3	3.5	6.24	5.25	0	0
-0.2	12.48	4.97	3.75	1.31	0
-0.1	9.56	4.29	5.7	4.51	7.02
0	8.09	5.24	5.29	5.48	4.4
0.1	8.18	4.7	5.27	5.37	4.97
0.2	9.7	5.86	5.24	5.24	5.06
0.3	8.75	5.26	4.09	4.73	5.14
0.4	8.82	5.16	5.14	5.96	4.87
0.5	6.55	4.71	6.48	4.79	6.21

Additionally, when analyzing the dataset, we can observe that the distribution of users along the different values for the five personality traits (-0.5 to 0.5) is unbalanced. Figure 1 shows the way in which the users are distributed in the dataset. The personality traits are located on the x-axis of the graph, while the values of the traits are located on the y-axis, the colors indicate the number of users in for each value for each trait, the red color color indicates more users and the blue color less users. From the figure we observe that most of the users fall in the middle values for all the traits, indicating that most of the subjects are functional from a psychological point of view. We also observe there is no presence of subjects without a trait, and that the trait agreeableness is present with a higher degree than other traits.

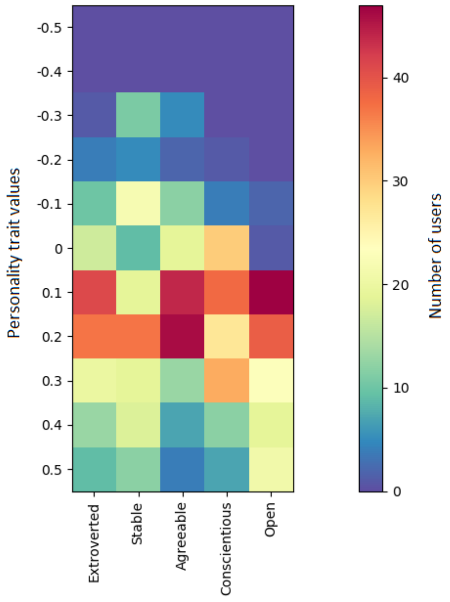

Fig. 1. PAN CLEF 2015 dataset: distribution by personality traits

In this work, similar to other works in the literature [10], we tackle the prediction of personality traits as a classification problem, where each observed value for a trait is considered a class. So, if for a personality trait we observe only three values (0.1, 0.0, -0.1), then we implement a classifier with three classes. Table II shows the range of the observed values and the number of classes for each personality trait in the PAN CLEF 2015 dataset.

TABLE II
PAN CLEF 2015 DATASET: RANGES OF ACTUAL VALUES AND NUMBER OF CLASSES PER PERSONALITY TRAIT

Trait	Range	Classes
Extroverted	[-0.3,0.5]	9
Stable	[-0.3,0.5]	9
Agreeable	[-0.3,0.5]	9
Conscientious	[-0.2,0.5]	8
Open	[-0.1,0.5]	7

B. Personality prediction model

Before building a classification model, we preprocess the dataset with several standard techniques. We first concatenated the text from all the tweets corresponding to each user in a single long document. We then split each document into five different content-based features: words, links, emojis/emoticons, mentions (ats) and tags (hashtags). Finally, we removed punctuation marks, very short words (length < 3), very long words (length > 35) and stopwords using the list of English stopwords provided by the Python NLTK library. For this work we focus on words and emojis/emoticons as feature sets, since they represent the majority of relevant content in the data.

After cleaning the text, we transform the document to a matrix representation using the TF-IDF weighting scheme. The TF-IDF transform is the product of two measures, frequency of term and inverse document frequency. TF-IDF is one of the most versatile statistic that shows the relative importance of a word or phrase in a document or a set of documents in comparison to the rest of a corpus, and is one of the most used ways to represent documents in text mining. Zhang et al. [18] show the classical equation of TFI-DF used to weight the relevance of a term (word) as:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

where $w_{i,j}$ is the weight for term i in document j, N is the number of documents in the collection, $tf_{i,j}$ is the term frequency of term i in document j and df_i is the document frequency of term i in the collection.

After building the TF-IDF matrix, we apply over the different techniques of dimensionality reduction mentioned before: PCA, LDA and NMF. With these techniques we sought to extract latent features to improve the classification for the detection of personality traits; considering that the latent features would represent the original data in a better more discriminative space.

Richardson [19] explains that the PCA method uses a vector space transform in order to reduce the dimensionality of large datasets. Applying a mathematical projection, the original dataset that possibly has many variables, can often be represented in a few variables (the main components). Having a data set with fewer dimensions makes it easier to detect patterns or trends in the data. The PCA constructs a linear transformation that selects a new coordinate system for the original dataset in which the largest variance of the dataset is captured on the first axis (called the First Major Component), the second largest variance is the second axis, and so on. To construct this linear transformation, the covariance matrix or matrix of correlation coefficients must first be constructed. Due to the symmetry of this matrix there is a complete basis of eigenvectors of it. The process to transforms from the old coordinates to the coordinates of the new base is precisely the linear transformation necessary to reduce the data dimensionality. In addition, the coordinates in the new base give the composition in underlying factors of the initial data.

The LDA technique consists of projecting the original data matrix into a space with fewer dimensions. The process of this technique is divided into three steps. The first step is to calculate the separability between different classes (that is, the distance between the means of different classes), which is called variance between classes. The second step is to calculate the distance between the mean and the elements of each class, which is called the variance within the class. The third step is to construct the space of minimum dimensions that maximizes the variance between classes and minimizes the variance within classes [20].

The Non-Negative Matrix Factorization (NMF) is a technique whose main utility is to find a linear representation of the data that has to be non-negative. Two of the main properties of the NMF method is that it usually provides a sparse representation of the data, and that the result is fairly simple to interpret. In this method, we decompose a data matrix V (in our case the TF-IDF matrix with the documents) in two matrices W and H.

$$V \approx WH \tag{2}$$

where W is a matrix of dimensions nxr that contain the vectors of the bases in their columns. The columns of H contain the coefficient vectors corresponding to the combinations of the vectors of the base that generate the measurement vectors. The matrix H is known as the coding matrix.

LDA is based on extracting discriminative features among classes, while PCA and NMF are based on extracting descriptive features of the whole dataset.

To evaluate the performance of the different latent features for predicting personality traits, we conduct 10-fold cross-validations over the PAN CLEF 2015 dataset. We experimented with three different classification methods: linear support vector classifier (LSVC), logistic regression (LR) and random forest (RF). We implemented the preprocessing, the dimensionality reduction methods and the classification methods in Python, using the libraries NLTK, numpy and scikit-learn [2]. We used the default parameters from the scikit-learn for each classification method.

When doing the 10-fold cross-validation, for every iteration of the validation we built an independent TF-IDF matrix using 9 folds, we applied the dimensionality reduction techniques to transform it, and then used the extracted vocabulary from the training part and the transformation matrices form the dimensionality reduction to transform the test fold.

IV. Experimental results

In this section we analyze the different dimensionality reduction techniques to extract latent features from textual information and their performances for predicting personality traits. We used as baseline the TF-IDF features, and we compare it with the use of the methods: PCA, LDA and NMF. We are interested in observing the contribution of the latent features that are extracted from these techniques for the detection of personality traits. Due to the small dataset we have worked, the imbalance in the data and the number of classes we consider (one class for each value), we recommend to take the results with caution. A correct or incorrect prediction is enough to change the results considerably.

Table III shows the accuracy results for the baseline and the three dimensionality reduction methods using LSVC. In this table we observe that for the extroverted and open traits, TF-IDF obtains better results than the other methods; while for the traits of stable, agreeable and conscientious, LDA technique presents better results. PCA only stands out in the prediction of the extraversion trait, equaling in performance to the TF-IDF technique. With the NMF technique no outstanding results were obtained. On average, the best performance comes from the LDA method. That means that the latent features extracted with it contain better discriminative power than other features.

TABLE III
LSVC: Classification accuracy for predicting personality traits

Trait	TF-IDF	LDA	PCA	NMF
Extroverted	**0.7**	0.67	**0.7**	0.38
Stable	0.65	**0.67**	0.64	0.41
Agreeable	0.67	**0.73**	0.67	0.34
Conscientious	0.63	**0.73**	0.63	0.37
Open	**0.73**	0.69	0.72	0.46
Average	0.68	**0.70**	0.67	0.39

Table IV and Table V show the results for the TF-IDF and the dimensionality reduction methods using logistic regression (LR) and random forest (RF) methods, respectively. Comparing the results with the ones shown in Table III, it is clear that LSVC presents better results when compared with the other two classification methods.

Finally, Table VI shows the results presented in [10], the winners of the PAN CLEF 2015 AP task. In this work, the authors represent the documents using a TF-IDF schema and

[2] Code available at: https://github.com/jcgcugto/personality_traits_prediction_conielecomp_2019

TABLE IV
LR: CLASSIFICATION ACCURACY FOR PREDICTING PERSONALITY TRAITS

Trait	TF-IDF	LDA	PCA	NMF
Extroverted	0.55	0.38	0.55	0.31
Stable	0.5	0.32	0.5	0.31
Agreeable	0.49	0.43	0.49	0.3
Conscientious	0.59	0.42	0.59	0.28
Open	0.6	0.47	0.6	0.37
Average	0.55	0.40	0.55	0.31

TABLE V
RF: CLASSIFICATION ACCURACY FOR PREDICTING PERSONALITY TRAITS

Trait	TF-IDF	LDA	PCA	NMF
Extroverted	0.25	0.31	0.3	0.31
Stable	0.32	0.27	0.32	0.24
Agreeable	0.34	0.39	0.38	0.31
Conscientious	0.27	0.37	0.33	0.28
Open	0.36	0.39	0.35	0.33
Average	0.31	0.35	0.34	0.29

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY FOR PREDICTING PERSONALITY TRAITS BY [10] AND OUR METHOD

Trait	LSA+SOA	LDA+LSVC
Extroverted	0.87	0.7
Stable	0.85	0.65
Agreeable	0.8	0.67
Conscientious	0.78	0.63
Open	0.86	0.73
Average	0.83	0.70

then combine two dimensionality reduction techniques, LSA and SOA (Second Order Attributes) to obtain their final feature set, and then perform a 10-fold cross-validation. Their results as presented in their paper are better than the ones presented in this work; nevertheless, their results are not completely comparable with ours, because in their methodology, they first transform the whole corpus with TF-IDF, apply LSA and then use the cross-validation. The problem with this schema is that some words from the test part of each iteration of the validation is mixed with the training part, which increases the chance of correctly classify the test examples. In our case, we independently create the TF-IDF and then dimensionality reduction method for each iteration of the validation.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an study of different dimensionality reduction techniques to extract latent features from the textual content of tweets for predicting personality traits of Twitter users. We tested: PCA, LDA and NMF. We found that the latent characteristics of the LDA technique are very complete, because it obtains the best results in the prediction of three personality traits and in the two remaining traits presents acceptable results. In the case of the PCA technique, the results are very similar to those of TF-IDF, and we can consider it does not provide an improvement in the prediction of personality traits. Finally, NMF presents the lowest results. As general remark, the discriminative features extracted with LDA are more consistent for the task, than the descriptive features extracted with the other methods.

Future research directions to improve the proposed approach may focus on applying combinations of descriptive characteristics with latent characteristics to analyze whether the results improve. Other directions include to train independent classifiers with each set of features and combine their output; and the use of other more complex dimensionality reduction techniques to improve the separability between classes [21], [22].

ACKNOWLEDGMENTS

This research was supported by Fondo Sectorial CONACYT-INEGI, proyecto número 290910, and by project PRODEP-SEP folio UGTO-PTC-595, oficio 511-6/17-8074.

REFERENCES

[1] D. J. Ozer and V. Benet-Martínez, “Personality and the Prediction of Consequential Outcomes,” *Annual Review of Psychology*, vol. 57, no. 1, pp. 401–421, 2006.

[2] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, “Manifestations of user personality in website choice and behaviour on online social networks,” *Machine Learning*, vol. 95, no. 3, pp. 357–380, 2014.

[3] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[4] L. R. Goldberg, O. P. John, H. Kaiser, K. Lanning, and D. Peabody, “An Alternative ”Description of Personality”: The Big-Five Factor Structure,” *Journal of personality and social psychology*, vol. 59, no. 6, pp. 1216–1229, 1990.

[5] R. Lambiotte and M. Kosinski, “Tracking the digital footprints of personality,” *Proceedings of the IEEE*, vol. 102, no. 12, pp. 1934–1939, 2014.

[6] W. Youyou, M. Kosinski, and D. Stillwell, “Computer-based personality judgments are more accurate than those made by humans,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.

[7] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-*, vol. 23, no. 3, pp. 321–346, 2003.

[8] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, “Automatically profiling the author of an anonymous text,” *Communications of the ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[9] O. M. ulea and D. Dichiu, “Automatic profiling of Twitter users based on their tweets,” *CEUR Workshop Proceedings*, vol. 1391, 2015.

[10] M. A. Álvarez-carmona, A. P. López-monroy, M. Montes-y gómez, L. Villaseñor-pineda, and H. J. Escalante, “Inaoe's participation at pan'15: Author profiling tasknotebook for pan at clef 2015,” *In [Rangel et al. 2015]*, 2015.

[11] I. Pervaz, I. Ameer, A. Sittar, and R. M. A. Nawab, “Identification of author personality traits using stylistic features,” *CEUR Workshop Proceedings*, vol. 1391, 2015.

[12] A. Grivas, A. Krithara, and G. Giannakopoulos, “Author profiling using stylometric and structural feature groupings,” *CEUR Workshop Proceedings*, vol. 1391, 2015.

[13] C. Gonzlez-Gallardo, J. Torres-Moreno, A. Montes-Rendn, and G. Sierra, “Perfilado de autor multilinge en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales,” *Linguamtica*, vol. 8, no. 1, p. 2129, 2016.

[14] E. Stamatatos, “Ensemble-based author identification using character n-grams,” *CEUR Workshop Proceedings*, pp. 41–46, 2006.

[15] J. E. Lantz, “A survey of modern preaching,” *Quarterly Journal of Speech*, vol. 29, no. 2, pp. 167–172, 1943.

[16] A. Poulston, M. Stevenson, and K. Bontcheva, “Topic Models and ngram Language Models for Author Profiling Notebook for PAN at CLEF 2015,” in *Proceedings of CLEF*, 2015.

[17] M. Arroju, A. Hassan, and G. Farnadi, “Age, gender and personality recognition using tweets in a multilingual setting,” *CEUR Workshop Proceedings*, vol. 1391, 2015.

[18] W. Zhang, T. Yoshida, and X. Tang, “A comparative study of TF*IDF, LSI and multi-words for text classification,” *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[19] I. T. Joliffe, “Principal Component,” *Principal Component Analysis SE - 7*, vol. 36, no. 4, p. 432, 2002.

[20] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” *AI Communications*, vol. 30, no. 2, pp. 169–190, 2017.

[21] J. C. Gomez, E. Boiy, and M.-F. Moens, “Highly discriminative statistical features for email classification,” *Knowledge and information systems*, vol. 31, no. 1, pp. 23–53, 2012.

[22] J. C. Gomez and M.-F. Moens, “Using biased discriminant analysis for email filtering,” in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 566–575, Springer, 2010.