Journal of Computer and Communications, 2023, 11, 75-88
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Supervised Learning Algorithm on Unstructured Documents for the Classification of Job Offers: Case of Cameroun

Fritz Sosso Makembe, Roger Atsa Etoundi, Hippolyte Tapamo

Department of Computer Science, Faculty of Science, University of Yaoundé 1, Yaoundé, Cameroon
Email: fritz-oswald.makembe@facsciences-uy1.cm

How to cite this paper: Makembe, F.S., Etoundi, R.A. and Tapamo, H. (2023) Supervised Learning Algorithm on Unstructured Documents for the Classification of Job Offers: Case of Cameroun. *Journal of Computer and Communications*, **11**, 75-88.
https://doi.org/10.4236/jcc.2023.112006

Received: January 20, 2023
Accepted: February 24, 2023
Published: February 27, 2023

Abstract

Nowadays, in data science, supervised learning algorithms are frequently used to perform text classification. However, African textual data, in general, have been studied very little using these methods. This article notes the particularity of the data and measures the level of precision of predictions of naive Bayes algorithms, decision tree, and SVM (Support Vector Machine) on a corpus of computer jobs taken on the internet. This is due to the data imbalance problem in machine learning. However, this problem essentially focuses on the distribution of the number of documents in each class or subclass. Here, we delve deeper into the problem to the word count distribution in a set of documents. The results are compared with those obtained on a set of French IT offers. It appears that the precision of the classification varies between 88% and 90% for French offers against 67%, at most, for Cameroonian offers. The contribution of this study is twofold. Indeed, it clearly shows that, in a similar job category, job offers on the internet in Cameroon are more unstructured compared to those available in France, for example. Moreover, it makes it possible to emit a strong hypothesis according to which sets of texts having a symmetrical distribution of the number of words obtain better results with supervised learning algorithms.

Keywords

Job Offer, Underemployment, Text Classification, Imbalanced Data, Symmetric Word Distribution, Supervised Learning

1. Introduction

In 2020, according to the World Bank [1], the unemployment rate in Cameroon was estimated at 3.4% and the underemployment rate was estimated at 84.7%. According to the International Labor Organization ILO [2], underemployment

occurs when the duration or productivity of a person's employment is inadequate in relation to other possible jobs that the person is willing and able to do. In Cameroon, it is interpreted as a failure of the labor market, and is characterized mainly by the misuse of professional skills. Moreover, with the advent of ICT (Information and Communication Technologies), we are witnessing a new form of labor market in Cameroon that of the supply and demand of jobs online from websites or other mobile applications. It is in this sense that Jonas Hjort *et al.* (2019) [3], provide evidence of the impact of the internet on the labor market in 12 African countries. Farrukh Suvankulov *et al.* (2012) [4] show that job seekers who used the internet saw their probability of being reemployed within 12 months increase from 7.1% to 12.7%. Good communication in this market requires the categorization of job offers. Indeed, job boards group job offers into categories, but in Cameroon, these categories correspond to the company's field of activity and not to the description of the offer itself. Thus, the failure of the job market is born because we have offers that are poorly classified from the point of view of the job seeker. It is therefore necessary to be able to categorize these offers so that the applicant can easily find the offers that best correspond to his or her profile. R. Feldman and J. Sanger [5] define automatic text classification as the task of classifying a data instance into a predefined set of categories, *i.e.*, given a set of categories (topics, classes, or labels) and a collection of text documents, classification is the process of automatically identifying the correct topic (or topics) for each document. So we can define job classification as the process of automatically putting together job offers that are similar. In other words, it can be compared to the automatic detection of the field to which a job offer relates according to its content. The objective of text classification is therefore to automatically classify documents into categories that have been defined either beforehand by an expert or automatically. This is supervised classification when the labeling is done by an expert and unsupervised classification (or clustering) when the labeling is done automatically by a machine. The rest of our work will focus mainly on supervised classification.

In our context, IT offers, for example, include offers from frontend developers, backend developers, community managers, database managers, assistants in an internet cafe or trainers on the office pack. The computer scientist is the one who uses the computer to solve a problem. Many works address the problem of supervised classification of job offers, and propose several approaches to solve it. However, our experiments have shown that, the approaches proposed in the literature have been shown to be insufficient for Cameroonian jobs. Indeed, the naive bayes, decision trees, SVM and recurrent neural networks methods give less good results on Cameroonian offers. The question then arises as to why these approaches or classical methods of classifying job offers, proposed in the literature, provide less good results on Cameroonian job offers.

To answer this question, we use data retrieved from the websites Minajobs (in Cameroon) and Monster (in France), which specialize in the publication of job offers on the Internet. These data allowed us to have two corpora of documents

(one corpus for each site). The study, using supervised learning algorithms, explores the impact on the precision of these algorithms by analyzing the distribution of the number of words in the documents of the corpus. That is to say that we wish to show, in an experimental way, that the distribution of the number of words, in a corpus, must be symmetrical to reach the optimal results with the above mentioned algorithms.

The rest of this work is divided into five main parts: a presentation of related work, followed by a presentation of the data and methods used, then a presentation of the results obtained, after a discussion and finally a conclusion.

2. Related Works

Many works have approached the classification of texts in general and job offers in particular in several ways and with different methods.

In attempting to improve the results of recommender systems by matching job offers and profiles according to required skills and experiences, A. Casagrande *et al.* (2017) [6] following the work of Dieng (2016) [7] and Florea *et al.* (2013) [8] [9], propose to automatically detect the sector of activity of the job offers using supervised learning techniques. The idea is to automatically assign a class (a universe or sector of activities) to a document (a job offer).

Moldagulova *et al.* (2017) [10] propose an approach for building a machine learning system in R that uses the KNN method for text document classification. Moreover, they show that: the impact of the value of k (which represents the number of neighbors) on the classification accuracy in the K-Nearest Neighbors (KNN) algorithm, is less from a high number of K. They first start by analyzing the text of two articles collected from two sites: (egov.kz; http://www.government.kz/). Then they use the word cloud technique to select the most frequent terms in the text; then they modify the documents into a more manageable representation: a vector of terms and their frequencies represents a document; finally they deploy the KNN algorithm by training the model with "known" data and then classifying it on "unknown" data. The model presented in this article has two main limitations: first, the choice of the parameter K, although having demonstrated that its impact on the accuracy of the classification decreases when it is large, there remains the problem of determining the optimal value of K which varies according to the corpus. Secondly, in this classification method, the model is the entire training corpus, which poses the problem of complexity in time and space because it is necessary to load the entire training corpus and recomputed the similarity with all the elements of the corpus when we wish to classify a new element. This second limitation poses a real problem on our corpus with large job offers.

Ouchiha, L. (2016) [11] having made a comparative study of supervised text classification methods, it emerges according to his study that SVM stands out and occupies the first place by its performance. Despite the fact that the performance of polynomial SVM far exceeds that of the decision tree (DA), we note that its execution time is significantly greater than that of the DA. This state of affairs led us to use the linear SVM available on WEKA, which gave very good

performances both in terms of classification error rate and execution time. He also demonstrated that the Naive Bayes Classifier (NBC) also performs well with long documents, due to its particular implementation in KNIME, as in the case of SVM. He makes the following remark, as he adds categories, the performance of AD deteriorates more and more, the interpretation which seems logical, is the fact that our AD is subjected to a very large dimension of descriptors which led to its over-learning.

Kameni F. *et al* (2020) [12] are interested in the extraction of skills expressed in documents such as CVs or job offers and based on the CNN (convolutional neural network) classification model manage to extract high level skills in CVs with performances reaching 98.79% for recall and 91.34% for precision. However, these data are retrieved in a very formal context.

Jakub Nowak *et al.* (2020) [13] address, using the supervised methods, the problem of non-uniformity of job names and descriptions by proposing two models: a convolutional network for text classification, consisting of six convolutional layers and three fully connected layers, and a recurrent network with long-term memory (LSTM) and Gated Recurrent Unit (GRU) cells with a convolutional input layer. In this solution, the description of the offer is entered word by word in the order in which it is written, this procedure simulates reading an ad on the Internet in the same way as humans. The convolutional part encodes the written word for the purposes of the recurrent cells, and provides an input vector to the output of the convolutional part. Therefore, all feature maps are combined into a single dimension given to the recurrent cells. The final classification remains with the LSTMs and GRUs. The number of calls to the recurrent cells was dynamic and depended on the number of words for each case in the database. The limitation was placed on the number of letters in a word and was 16 characters. Also, the number of words per offer was limited to 1024 and for any offers exceeding this limitation, the words after the 1024th were not taken into account. In addition to these adjustments, the SELU activation function was used throughout the framework as an alternative to the widely used RELU function, and they justify this by the fact that the SELU function can give negative values, which speeds up the learning process of the convolutional network. They applied their models on 17,177 job offers obtained from the Emplocity Ltd website (https://emplocity.com/) grouped into five classes. They obtain an accuracy of 84.7% for the LSTM and 86.5% for the GRU.

The various works mentioned above study different aspects of job postings using supervised learning methods, but none of them focuses on Cameroonian job postings. This paper aims to demonstrate the particularity of Cameroonian jobs offers and to measure the precision of naive bayes, SVM, and decision trees algorithms on a corpus of Cameroonian computer job offers taken from the internet. The results will be compared to those obtained on a set of computer job offers obtained on the Monster website in France.

To do so, we will start by submitting the offers to a new pre-processing approach aiming at normalizing our offers. This pre-processing approach consists,

firstly, in removing the job offers with atypical word counts up to a certain threshold. In fact, by removing atypical offers, we should be able to keep a representative number of starting offers, and the remaining atypical offers should be negligible or even zero. Secondly, we apply tokenization and stemming, and finally we remove stopwords and other special characters. Then the offers that have been subjected to the new preprocessing approach are transformed into frequency vectors with the TF-IDF method as done by Diaby, M *et al.* (2014) [14] [15]. Finally on these offers transformed into frequency vectors, we apply supervised learning methods to classify them and evaluate the classification results.

3. Data and Methods

3.1. Data

We have 7533 job offers from minajobs.net [16] in all sectors of activity. However we concentrated on the offers in French and in computer science. This choice was made because we were not able to obtain the labeling of the job offers by the experts of the other fields. We believe that this choice does not impact the results in the other categories because job offers in project management or marketing encounter the same difficulties. However, we have some reservations about areas such as medicine or academic training.

Thus, as shown in **Figure 1** below, we have a corpus of 726 job offers distributed over 13 classes numbered from 1 to 13 and corresponding respectively to the categories: developer, web master, database manager, analyst, digital marketing, designer, network administrator, IS security, computer maintenance, community manager, archivist, tester, system administrator.

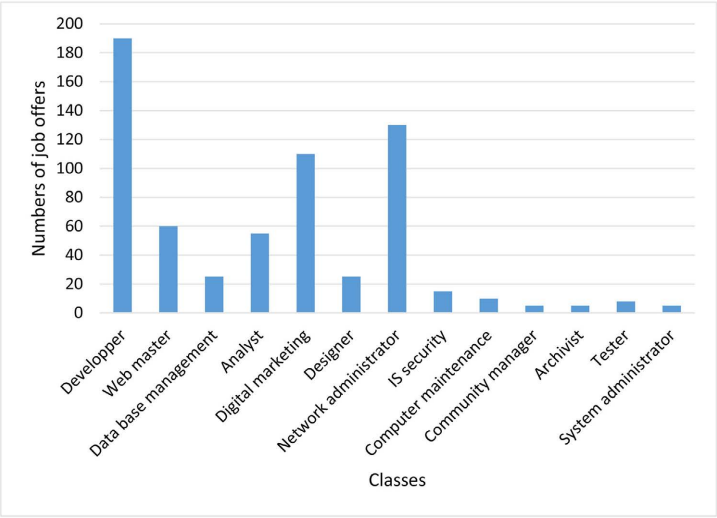

Figure 1. Distribution of the number of offers by classes.

This corpus has offers belonging to the same category (or domain) but whose lengths (*i.e.* the number of words contained in the text of the offer) are very different. This is the case of the example presented in the following images. These are two offers belonging to the "developer" domain, one has three words and the

other has 2341 words, so a difference of 2339 words. In **Figure 2**, anyone who has done computer development can apply. The candidate knows nothing more about the client's need. In **Figure 3**, the offer has a part in English because Cameroon is bilingual; however, we have used a language detection algorithm which classifies it in French because the majority of the text is in French. Several jobs are available in the same offer.

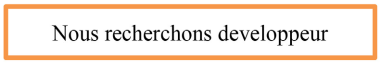

Figure 2. Job offer taken on minajobs with only three words (source: minajobs).

MINISTERE DE L'ECONOMIE, DE LA PLANIFICATION ET DE L'AMENAGEMENT DU TERRITOIRE PROGRAMME NATIONAL DE DEVELOPPEMENT PARTICIPATIF (PNDP) COMMISSION SPECIALE DE PASSATION DES MARCHES (CSPM) Manifestation d'intérêt Sollicitation à manifestation d'intérêt N°008 pour la sélection d'un Consultant chargé de la finalisation de
……
…… résidence de l'ambassadeur de Côte d'ivoire, non loin des bureaux de l'Organisation des Nations Unis pour l'Education…Le Coordonnateur National Marie

Figure 3. Job offer taken on minajobs of 2341 words (source: minajobs).

The second dataset is a set of job offers collected on the French website Monster.fr.

The offers used in [15] are not accessible, because to have access to the platform on which the offers were extracted, it is necessary to make a physical and paying registration. We have therefore retained the French offers because of their accessibility and because, in terms of structure, they are closer to ours than to the offers used in [15]. As shown in **Figure 4**, it is a corpus of 1280 job offers, in data processing, opened on Monster.fr and divided into three classes numbered from 1 to 3 corresponding, respectively, to the categories: computer sciences engineer, computer graphics and software architect. Now we will explain the methodology for comparing the performance of the supervised classification algorithms on the two corpora.

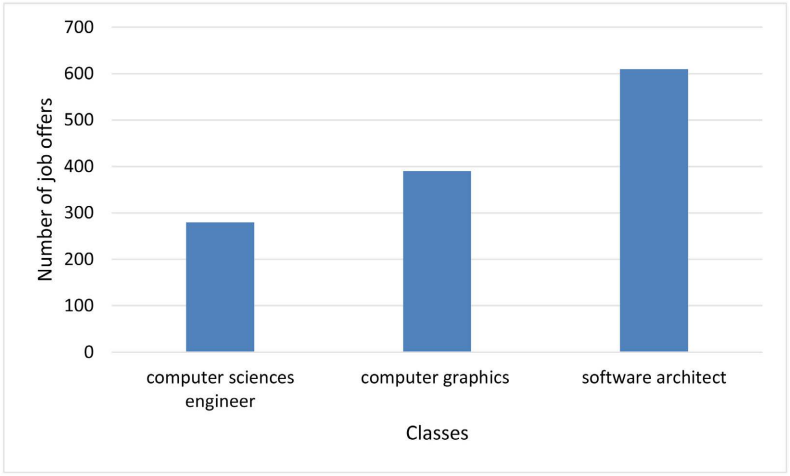

Figure 4. Distribution of the number of offers by classes on monster.fr.

3.2. Methods

The methodological approach used in this work is presented in Figure 5. The first step is the cleaning step which consists in removing the atypical descriptions, *i.e.* those whose word count contrasts greatly with the "normal" measured values. Indeed, as Yamada *et al.* (2020) [17], we based ourselves on the measure of the interquartile range to determine the atypical offers. Thus, all offers outside the interval: [Q1 − k(Q3 − Q1), Q3 + k(Q3 − Q1), where k is a positive constant, Q1 and Q3 are the first and third quartile respectively. Then we cut the descriptions into lists of words (tokenization) to transform these words into their root or radical, also called the stems (stemming). Finally, we remove the stopwords.

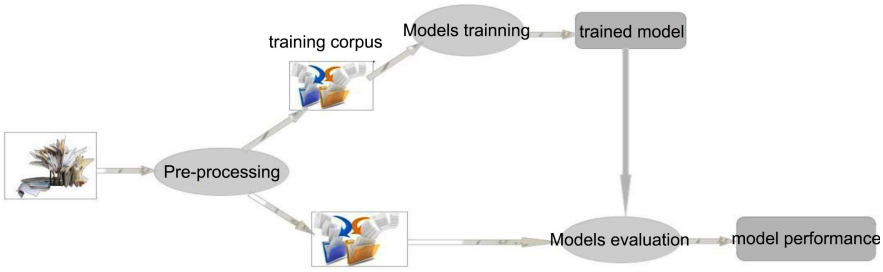

Figure 5. Methodological approach for the classification of job offers (source: author).

3.3. Pre-Processing of Job Offers

The preprocessing approach (Figure 6) is shown in the following diagram:

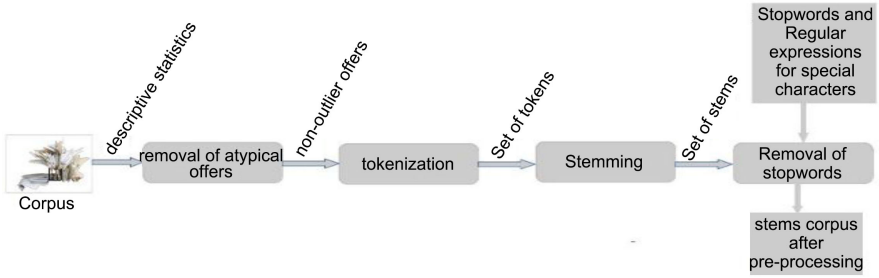

Figure 6. Methodological approach for the preprocessing of job offers (source: author).

Step 1) Removal of outliers: In this step, we remove the outliers, *i.e.* the offers with an atypical number of words. To do this, we are mainly interested in the descriptive statistics of the series of numbers of words per offer, from which we determine the offers that have atypical numbers of words (visible on the whisker box) and we remove them. The goal here is to obtain a histogram of the distribution of Cameroonian offers that is as close as possible to a symmetrical distribution.

Step 2) Tokenization: According to Wisdom *et al.* (1999) [18] tokenization consists in transforming a text into a list of words without separators. In our case it is to separate an offer into a list of words

Step 3) Stemming: According to Perkins *et al.* (2010) [19] stemming is a method which consists in extracting the roots of the words. In our case, for each

word obtained after tokenization, we will apply stemming and obtain a new list of words that will be the radical of the tokens, this radical is still called stem.

Step 4) Removal of stopword: Here we are going to remove words that are devoid of information. For this purpose, we have developed a stopword dictionary; in addition, regular expressions are used to remove certain elements such as special characters.

Step 5) Vector representation of job offers: In this step, we numerically represent the job offers using the TF-IDF method. Each job offer being already a stem list will be represented by a vector.

$$TF\_IDF = TF * IDF \quad [20] \tag{1}$$

Where:

$$TF = \text{Number Of Stems Occurrences} / \text{Number Of Stems} \quad [20] \tag{2}$$

$$IDF = \log\left(\text{Number Of Descriptions} / \text{Descriptions Containing Stem}\right) \quad [20] \tag{3}$$

Step 6) Supervised classification methods: We are doing single-label classification with these three supervised classification algorithms:

- Support Vector Machine (SVM): We used the linear SVM.
- Naïve Bayes: We thus obtain the naive bayes Gaussian.
- Decision tree: We used the classification model based on the ID3 [21] algorithm

4. Results

Here we present the results of our experiments. We present the results of the application of the descriptive statistics of the data and the classification methods described in section 3.2.3 first on the two datasets that did not undergo any pre-processing beforehand, then on the same data but this time after deleting all the offers with an atypical number of words, and finally on the two datasets after having applied all the pre-processing approach described in section 3.3.

4.1. Results without the Proposed Approach

After the statistical study carried out on our different corpora without having previously applied any pre-processing, we obtain the following results (**Table 1**).

From these observations, we notice that the Monster offers have two atypical offers while the Cameroonian offers have 49; and especially that the difference between the extents of the two corpora is 1520. In addition, the histograms (**Figure 7**, **Figure 8**) show that the distribution of the number of words of the Cameroonian offers is very spread on the right compared to the distribution of words of the Monster corpus. This is also justified by the empirical skewness values obtained.

After supervised learning on the corpora, the performances obtained by the different classification methods have been summarized in the following table with precision as the performance evaluation metric (**Table 2**).

The previous table shows the difference in performance of the classification

methods on our corpus. Indeed, overall, these performances are clearly better on Monster's offers. This can be justified by the difference in the shape of the distribution between the two corpora.

Table 1. Comparison of descriptive statistical study done on Monster and Minajob job postings (source: Author).

Elements of comparison	Offers from Minajobs	Offers from Monster
Total number of words	227,113	22,817
Average number of words	312.828	334.538
Minimum number of word	3	64
Max number of word	2341	885
Median	200	333
Standard deviation	257.591	126.969
Number of atypical values	49	2
Scope	2338	821
Empirical skewness	2.531	0.356

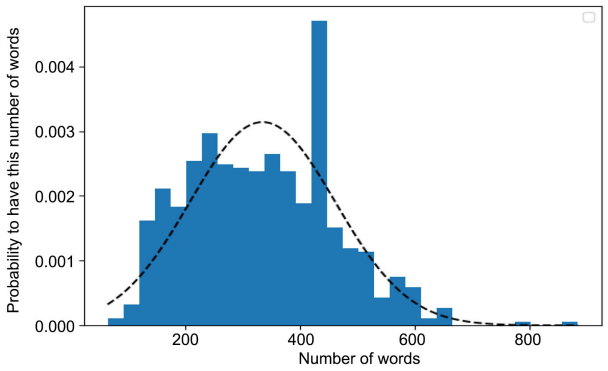

Figure 7. Histogram of the distribution of the number of words in Monster's offers.

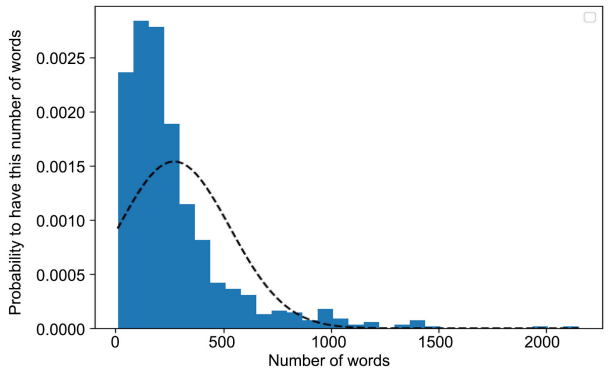

Figure 8. Histogram of the distribution of the number of words in Minajob offers.

Table 2. Results obtained by supervised learning with Naive Bayes, decision tree, and SVMmethods on Monster and Minajob job offers without preprocessing (source: author).

Methods	Naïve Bayes		Decision tree		SVM	
Data Metrics	Minajobs offers	Monster offers	Minajobs offers	Monster offers	Minajobs offers	Monster offers
Recall	62.79%	85.99%	69.47%	86.56%	69.34%	86.45%
Precision	62.79%	86.04%	68.02%	87.37%	68.27%	87.86%
F1-score	59.79%	85.49%	65.98%	86.09%	65.88%	86.49%

4.2. Results after Deleting Offers with an Outlier Number of Words

In this part, we first removed all the offers with an atypical number of words (the outliers), then we redid the statistical study on the two corpora without outliers, and finally we performed the classification again. After the statistical study carried out on our different corpora without outliers, the result is summarized in the following Table 3.

The following figures present respectively the histograms of the distributions of the number of words of the Cameroonian offers after removing the outliers.

After removing the offers with an atypical number of words, we see, in Figure 9, that the range of the distribution of words in the Cameroonian offers has decreased by 1635. The maximum number of words in an offer is now 706, the skewness has also decreased by 1.884651. Thus, the curve and the histogram of the distribution of words in the Cameroonian offers are closer to those of the Monster offers. On the other hand, the Monster offers have fewer changes.

After supervised learning on the corpora after having eliminated the outliers, the performances obtained by the different classification methods have been summarized in the following table with precision as the performance evaluation metric (Table 4).

The previous table shows us the difference in performance of the classification methods on our corpora after the removal of outliers. We can see that the results obtained on the Cameroonian corpus have considerably increased with the removal of outliers. The performances are now close to those obtained on the Monsters offers which have not changed significantly. This can be justified by the distribution of words in the corpora, indeed by removing the outliers on the Cameroonian corpus; we have considerably modified the distribution of words in this corpus, giving it a distribution close to that of the Monster corpus. However, this deletion did not have a major influence on the offers of the Monster corpus because it contained only two outliers.

4.3. Results after Complete Pre-Processing of Offers

Having already the corpora devoid of outliers, we continued the pre-processing with tokennization, stemming, and stopwords removal on the offers. Thus, after finishing with the pre-processing we redid supervised learning on the pre-processed corpora. The performances obtained by the different classification methods have

been summarized in the following table with precision as the performance evaluation metric (Table 5).

The previous table allows us to see that the performance of the classification methods on the two corpora has increased by an average of 9% for the corpus of Cameroonian offers and by 7% for the Monster offers. This implies that the tokennization, stemming and stopwords removal steps have improved the classification of the offers. The fact that this increase is 2% more important for the Cameroonian offers than for the Monster offers allows us to say that the Cameroonian offers have a little more stopwords than the Monster offers.

Table 3. Comparison of statistical study done on Monster and minajobs job postings after removing outliers (source: author).

Elements of comparison	Offers from Minajobs	Offers from Monster
Total number of words	174,026	22,446
Average number of words	257.055	333.032
Minimum number of word	3	64
Max number of word	706	661
Median	181	333
Standard deviation	119.016	124.085
Number of atypical values	0	0
Scope	703	597
Empirical skewness	0.646	0.212

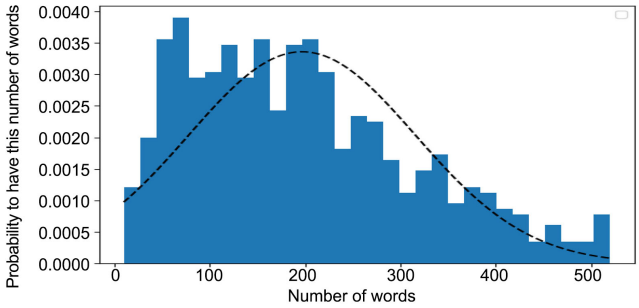

Figure 9. Histogram of the distribution of the number of words of the Cameroonian offers after deletion of the outliers (source: author).

Table 4. Results obtained by supervised learning with Naive Bayes, decision tree, and SVM methods on Monster and Minajob job offers after eliminating outliers (source: author).

Methods	Naïve Bayes		Decision tree		SVM	
Data / Metrics	Minajobs offers	Monster offers	Minajobs offers	Monster offers	Minajobs offers	Monster offers
Recall	93.89%	86.55%	86.75%	86.42%	86.97%	86.01%
Precision	94.93%	87.95%	87.75%	87.93%	87.97%	87.47%
F1-score	94.42%	86.55%	85.75%	86.89%	85.97%	86.18%

Table 5. Results obtained by supervised learning with Naive Bayes, decision tree, and SVM methods on Monster and Minajob job offers after complete pre-processing of these (source: author).

Methods	Naïve Bayes		Decision tree		SVM	
Data / Metrics	Minajobs offers	Monster offers	Minajobs offers	Monster offers	Minajobs offers	Monster offers
Recall	93.79%	90.88%	95.99%	96.79%	95.12%	96.23%
Precision	94.90%	88.95%	97.45%	97.56%	97.90%	97.96%
F1-score	93.75%	88.98%	96.65%	96.49%	96.90%	96.87%

5. Results Analysis

The different experiments carried out during this work allow us to see that the classic approaches to classifying job offers give less satisfactory results on Cameroonian job offers. On the other hand, when we change the distribution of the words of these offers by eliminating the offers having an aberrant length, which in our case constituted only 7% of the offers, we make more symmetrical the curve of distribution of the words of the offers and thus, increase by nearly 20% the precision of the classification methods on these offers. When, in addition to the removal of outliers, we add tokenization, stemming and stopword removal, we increase the precision of the classification methods by almost 9%. This allows us to conclude that the Cameroonian job offers have a main problem on their word count distribution. Indeed, Table 1 shows us that the Cameroonian corpus has a word count of 2338, meaning that some job offers have a very high word count compared to other offers in the same corpus, which generally biases the vector representation of the offers and therefore the classification.

6. Conclusion

This research also provides a strong hypothesis that text sets with a symmetric distribution of word counts are more likely to perform better with supervised learning algorithms. The results of the research indicate that, in the Cameroonian context, published offers must often be reprocessed to better match the expectations of employers and job seekers. In this context, we believe that the hypothesis could be justified by demonstrating why the results of job advertisement classification are better when the distribution of the number of words per advertisement is closer to a symmetric distribution and find out to what extent, instead of removing aberrant offers, they should be corrected, because even if they are aberrant, they are still offering to be considered.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Bank Website (2012) Cameroon: Universities Debate Unemployment.

https://www.worldbank.org/en/news/feature/2012/03/22/cameroon-universities-debate-unemployment

[2] International Definitions and Prospects of Underemployment Statistics (2021). https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_091440.pdf

[3] Hjort, J. and Poulsen, J. (2019) The Arrival of Fast Internet and Employment in Africa. *American Economic Review*, **109**, 1032-1079. https://doi.org/10.1257/aer.20161385

[4] Suvankulov, F., Lau, M.C.K. and Chau, F.H.C. (2012) Job Search on the Internet and Its Outcome. *Internet Research*, **22**, 298-317. https://doi.org/10.1108/10662241211235662

[5] Feldman, R. and Sanger, J. (2006) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511546914

[6] Casagrande, A., Gotti, F. and Lapalme, G. (2017) Classification d'offres d'emploi. University of Montreal, Montreal. https://rali.iro.umontreal.ca/rali/node/1519/

[7] Dieng M.A. (2016) Développement d'un système d'appariement pour l'e-recrutement. Université de Montréal, Montréal.

[8] Florea, N.V. (2013) Cost/Benefit Analysis—A Tool To Improve Recruitment, Selection and Employment in Organizations. *Management & Marketing*, **11**, 274-290.

[9] Pazzani, M.J. and Billsus, D. (2007) Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A. and Nejdl, W., Eds., *The Adaptive Web. Lecture Notes in Computer Science*, Vol. 4321, Springer, Berlin, 325-341. https://doi.org/10.1007/978-3-540-72079-9_10

[10] Moldagulova, A. and Sulaiman, R.B. (2017) Using KNN Algorithm for Classification of Textual Documents. 2017 8*th International Conference on Information Technology* (*ICIT*), Amman, 17-18 May 2017, 665-671. https://doi.org/10.1109/ICITECH.2017.8079924

[11] Ouchiha, L. (2016) Classification supervisée de documents: Étude comparative. Université du Québec en Outaouais, Gatineau. https://di.uqo.ca/id/eprint/806

[12] Jiechieu, K.F.F. and Tsopze, N. (2020) Skills Prediction Based on Multi-Label Resume Classification Using CNN with Model Predictions Explanation. *Neural Computing and Applications*, **33**, 5069-5087. https://doi.org/10.1007/s00521-020-05302-x

[13] Nowak, J., Milkowska, K., Scherer, M., Talun, A. and Korytkowski, M. (2020) Job Offer Analysis Using Convolutional and Recurrent Convolutional Networks. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R. and Zurada, J.M., Eds., *Artificial Intelligence and Soft Computing. ICAISC* 2020. *Lecture Notes in Computer Science*, Vol. 12416, Springer, Cham, 380-387. https://doi.org/10.1007/978-3-030-61534-5_34

[14] Diaby, M. and Viennet, E. (2014) Taxonomy-Based Job Recommender Systems on Facebook and LinkedIn Profiles. 2014 *IEEE Eighth International Conference on Research Challenges in Information Science* (*RCIS*), Marrakech, 28-30 May 2014, 1-6.https://doi.org/10.1109/RCIS.2014.6861048

[15] Quang, C.T. (2005) Classification automatique des textes vietnamiens Hanoi. Institut de la Francophonie pour l'informatique, Hanoi, Vietnam.

[16] Minajobs.net. https://cameroun.minajobs.net/

[17] Yamada, Y., Shinkawa, K. and Shimmei, K. (2020) Atypical Repetition in Daily Conversation on Different Days for Detecting Alzheimer Disease: Evaluation of Phone-Call Data from a Regular Monitoring Service. *JMIR Mental Health*, **7**, e16790. https://doi.org/10.2196/16790

[18] Wisdom, V. and Gupta, R. (2016) An Introduction to Twitter Data Analysis in Python. Artigence Inc., Bangalore.

[19] Perkins, J. (2010) Python Text Processing with NLTK 2.0 Cookbook. Packt Publishing, Birmingham.

[20] Yoo, J.Y. and Yang, D. (2015) Classification Scheme of Unstructured Text Document Using TF-IDF and Naïve Bayes Classifier. *Advanced Science and Technology Letters*, **111**, 263-266. https://doi.org/10.14257/astl.2015.111.50

[21] Brownlee, J. (2016) Master Machine Learning. Melbourne, Australia. https://datageneralist.files.wordpress.com/2018/03/master_machine_learning_algo_from_scratch.pdf