Moroccan Code-Mixing and a Comparative Study of the Best-Ranked Classifiers

Fadoua Mansouri, Sadiq Abdelalim

Abstract: ***The code-mixing called also code-switching is a linguistic term which indicates the passage from one language to another or from a dialect to a language and vice versa, by speakers who have more than one language in common. So to be able to understand and study a community, we need to understand the different means of communication that this community uses.***

The main purpose of our research is studying and analyzing the opinions of Moroccan Internet users. Those opinions are mainly posted on the web and social media, thus the major goal of this paper is to describe this new sociolect language or code-mixing that Moroccan internet users use, also we will present in this paper a comparative study of the best-ranked linguistic classifiers that we run on a sociolect corpus that we have built by extracting data from the web.

Index Terms: Linguistic classifiers, Moroccan code-mixing, Moroccan sociolect language, Naïve Bayes classifier, KNN, SVM.

I. INTRODUCTION

The classification and categorization of documents is the action of the automatic processing of natural languages, which is automatically classifying documentary data, usually coming from a corpus.

Document or Text classification is one of the important and typical tasks in supervised machine learning. Assigning categories to documents, which can be a web page, library book, comments posted on social media etc. has many applications like spam filtering, email routing, sentiment analysis, etc.

Our aim is to be able to classify opinions of Moroccans Internet users, and when doing so, we noticed that those users use a new trend of communication, a new language that combines more than one idiom it is what we call here the Moroccan sociolect language. Thus is this paper we present further details about this new language.

The rest of this paper is organized as follows: Section 2 is devoted to discuss the code-mixing in the daily spoken Moroccan or what we agreed to call the Moroccan sociolect language, section 3 is describing how we collected data to build our corpus, section 4 is about the processing of the Moroccan sociolect language, section 5 is a presentation of our approach in running a comparative study of the best-ranked classifiers using the corpus we have built, And section 6 is to conclude and give some perspectives.

II. CODE-MIXING IN THE DAILY SPOKEN MOROCCAN

From a sentence to the next, from Arabic to French to English and back to Arabic, this kind of code mixing has become widely used in the daily spoken Moroccan.

Code-mixing also called code-switching is a linguistic term, which indicates the passage from one language to another or from one dialect to another and vice versa, it is a widespread multifunctional characteristic of the speech of bilinguals in formal and informal settings [1]. It is a complete shift from one language to another, either for a word, a phrase or a whole sentence.

Given the history of the Moroccan country, we can notice that there are several languages and / or dialects that coexist in Morocco. So by analyzing the linguistic situation of this country, we can conclude that there are three major linguistic groups [2]:

- The Arabic language group. It consists of two subsystems: the first is Standard Arabic also called classical, academic, standard or modern Arabic, and the second subsystem is the Dialect Arabic, also known as “Moroccan Arabic” or “Darija”
- The Amazigh language group. It is materialized, on the (socio) linguistic level, by the different dialectal varieties. And according to A.Mabrour [3], these are, in fact, the only observable reality on the ground, but the reality contradicts this assertion because Amazigh is taught at the primary school, there are also TV channels dedicated to this language, and there is also an artistic production quite present notably in cinema, theater music and festivals.
- The group of foreign languages. It is represented mainly by the French language and, to a lesser extent, by the Spanish language in northern and southern Morocco. Also there is English language, which has recently started to gain significance in Morocco [4].

Revised Manuscript Received on June 19, 2019

Fadoua MANSOURI, SIM team of MISC Laboratory
Faculty of Science, University IBN TOFAIL, Kenitra, Morocco
Sadiq ABDELALIM, SIM team of MISC Laboratory
Faculty of Science, University IBN TOFAIL, Kenitra, Morocco

The presence of several languages and dialects in Morocco and the exchange [5] between these languages gave birth to a new language used by Moroccans in writing on the web and especially on social networks, this language that we agreed to call “the Moroccan sociolect language” is mainly characterized by the simultaneous use of, numbers, Latin script or figures and / or icons (emoticons) to speak in Arabic which is a Hamito-Semitic language yet having its proper spelling.

We have opted for the term "sociolect" [6] because it is the term that describes the best the situation that we are describing. So nowadays Moroccan Internet users use words like “3a2ila”, that means “family”, “9dim” to say “old”, “5ssar” to say “loose”, “5awi” to say “empty”, “far7an” to say “happy”, etc. So if we want to analyze, for example, the trend of the opinions of Moroccan Internet users and if we want relevant results that reflect the reality, then we must study this sociolect language, which is now almost totally used by the majority of Moroccan netizens.

III. CORPUS CREATION

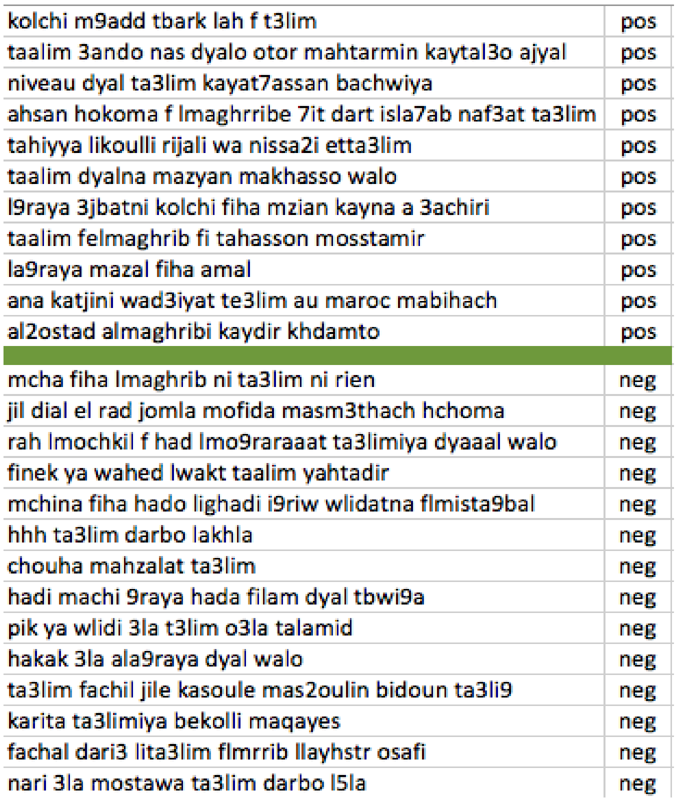
If Our final purpose is to be able to study the opinions posted by Moroccan net surfers on the web and namely on social media thus the next step were to use a corpus dedicated to this Moroccan sociolect language. Unfortunately, there is not yet an available corpus of opinion belonging to the Moroccan sociolect language, so we decided to build our own corpus, for that we chose a theme that is the state of education in Morocco.

We collected data from the Internet including Facebook pages most used by Moroccan Internet users since Facebook is most used social media in Morocco [7]. These pages were chosen According to Alexa Ranking [8], which provides a regular update of the most visited websites in Morocco and in other countries as well.

For data collection we used in addition to scraping solutions, a data extraction tool named Facepager [9] that was created to fetch public available data from Facebook, Twitter and other JSON-based API.

The data collected needed pre-processing and cleansing to make it workable, because the extracted comments did not all belong to the Moroccan sociolect language, so we needed to eliminate these sentences. In addition, we have manually eliminated sentences that are not complete, do not make sense and don’t belong to the chosen theme. Finally the data has been stored in a MySql database in order to apply different classifiers thereafter.

The corpus built is divided into two parts, the first is the training or learning set and the second is the test set. It should be noted that the training corpus has undergone a manual annotation, an annotation according to the polarity of the comment so we had an annotated part as a positive opinion and the rest as negative.



kolchi m9add tbark lah f t3lim	pos
taalim 3ando nas dyalo otor mahtarmin kaytal3o ajyal	pos
niveau dyal ta3lim kayat7assan bachwiya	pos
ahsan hokoma f lmaghrribe 7it dart isla7ab naf3at ta3lim	pos
tahiyya likoulli rijali wa nissa2i etta3lim	pos
taalim dyalna mazyan makhasso walo	pos
l9raya 3jbatni kolchi fiha mzian kayna a 3achiri	pos
taalim felmaghrib fi tahasson mosstamir	pos
la9raya mazal fiha amal	pos
ana katjini wad3iyat te3lim au maroc mabihach	pos
al2ostad almaghribi kaydir khdamto	pos
mcha fiha lmaghrib ni ta3lim ni rien	neg
jil dial el rad jomla mofida masm3thach hchoma	neg
rah lmochkil f had lmo9raraaat ta3limiya dyaaal walo	neg
finek ya wahed lwakt taalim yahtadir	neg
mchina fiha hado lighadi i9riw wlidatna flmista9bal	neg
hhh ta3lim darbo lakhla	neg
chouha mahzalat ta3lim	neg
hadi machi 9raya hada filam dyal tbwi9a	neg
pik ya wlidi 3la t3lim o3la talamid	neg
hakak 3la ala9raya dyal walo	neg
ta3lim fachil jile kasoule mas2oulin bidoun ta3li9	neg
karita ta3limiya bekolli maqayes	neg
fachal dari3 lita3lim flmrrib llayhstr osafi	neg
nari 3la mostawa ta3lim darbo l5la	neg

Figure 1. a catch of some annotated sentences belonging to the learning

IV. MOROCCAN SOCIOLECT LANGUAGE PRE-PROCESSING

Since this Moroccan sociolect language is a new trend of communication in Morocco, it is not recognized as an official language, so there are no rules and standards to follow in order to treat this new trend of language. Therefore we tried in previous works to introduce a new approach to recognize the Moroccan sociolect language [10] used on the web. This approach is found on identifying the language of each word that compose the sociolect text since this kind of text can be constructed by words that belong to different languages, for example the following sentence “ok, natla9aw 7da supermarché” that means “ok, lets meet near the market” contain the English word “ok”, the “natla9aw 7da” words that belong to the Moroccan sociolect language, and the “supermarché” that is a French word. So as we can notice there is more than one idiom in one sentence, thus this approach we proposed is based on using, for each language detected, a suitable dictionary. And as there is no available dictionary for the Moroccan sociolect language, we proposed, in another work [11], to build a first prototype of an electronic dictionary named DELSOM that is dedicated for this sociolect language. So in this work, we presented in detail the approach we followed in order to build this electronic dictionary, namely the general features of this knowledge base, the morphological and syntactic specifications, the different grammatical and phonetic rules, and the modeling schemes adopted to define the canonical form of the entries of this dictionary.

The following algorithm explain the steps we proposed in order to define the canonical form of a sociolect word, each time we have an input which is a sociolect word:

- *Detecting the grammatical category of the sociolect word (the category will be either a nominal or a verbal category)*
- *If the category is a nominal one, then we have two possible sub-categories: noun and adjective. And in this case we look for its masculine singular corresponding form.*
  - ➢ *There is one exception which is the broken plural sub category that we keep it as it is*
- *If the category is the verbal one, then we have two sub categories: verb and deverbale.*
  - ➢ *If we deal with a verb, then the canonical form corresponds to the third masculine person singular of the verb*
  - ➢ *If the word is a deverbale one, then we keep its masculine singular corresponding form*

The following algorithm explain the steps we followed to handle the phonetic rules adopted for the elaboration of the DELSOM dictionary entries and in this case we have the phonemes, composing the sociolect, word as input:

- *Detecting the grammatical category of the phoneme (the category will be either a vowel or a consonant)*
- *It the phoneme is a vowel and while the pronunciation of the vowel contains a vocal elongation then we repeat the vowel once*
- *It the phoneme is a vowel and while the pronunciation there is no vocal elongation then we don't repeat the vowel in the writing.*
- *If the phoneme is a consonant and while the pronunciation there is gemination so we repeat the consonant once in writing. But if there is no gemination we keep the consonant as it written once*
  - ➢ *When we have the letter "s" between two vowels we double it so as not to pronounce it "z".*

V. CLASSIFICATION OF MOROCCAN SOCIOLECT LANGUAGE

Since we are studying the opinions of Moroccan Internet users, the next step was to run a comparative study of the top-ranking classifiers to identify the most adapted one to the Moroccan sociolect language, and also to try to adjust some parameters to enhance their accuracy according to the feature that characterize this sociolect language.

A. *Naïve Bayes classifier*

a) Solution 1 (S1)

The first issue we faced when applying the Bayes theorem is the so call "the Underflow".

The Underflow [12] is a condition that occurs when arithmetic operations produce results too small to store in the available register. It is the opposite of overflow, which relates to a mathematical operation resulting in a number which is bigger than what the machine can store. To pass around this problem we have used the so-called additive smoothing, also called Laplace smoothing, which is a technique used to smooth categorical data. And this way we can handle now the probability of rare words that don't occur much in the text treated.
Let be the Laplace smoothing theorem:

$$P(w/c) = \frac{\mathrm{count}(w,c) + 1}{\mathrm{count}(c) + V} \qquad \square\ \square\ \square$$

Where:

- **count(W, C)** : Occurrence of the word W in documents belonging to category C
- **count(C)**: Number of words from documents belonging to category C
- **V**: Total number of words making up the learning corpus

After applying the Bayes classifier with, we got the following accuracy:

Accuracy (S1) = 69%

b) Solution 2 (S2)

For this solution we only considered the relevant words of the corpus, and we redid the calculation with the additive smoothing for the words whose occurrence is zero, that said that we eliminate articles, pronouns, etc.
The calculation of the accuracy for this approach gave the following value:

Accuracy (S2) = 62%

We notice that the first solution has a better accuracy compared to the second; however, this result is somehow illogical because it was expected that the second solution 'would be more precise because it is the solution where one gives more importance to the relevant words composing the commentary.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

After reflecting on this result, we can say that there is a reason behind it; is that the sociolect language is a little specific compared to other usual languages: The second solution does not take into consideration the collocations used in abundance by the sociolect language: for example, the expression "Mcha fiha" “مشا فيها” is a colocation meaning "lost", but when we take each word apart it changes meaning, it's not the same value anymore.

Also, in terms of relevant words, we will keep the term "mcha" “مشا” because only it is considered as a verb, and the verb counts a lot in a sentence while we will neglect the term "fiha" “فيها” as a preposition. So, a sentence containing all the collocation "mcha fiha" “ مشا فيها” is not at all close to a sentence containing only the word "mcha" “مشا” or the word "fiha" “فيها” and it is the error committed when we consider each word apart.

c) Solution 3 (S3)

For this third solution, we took into consideration the collocations specific to the sociolect language when choosing relevant word of the corpus, so we obtained the following results:

Accuracy (S3) = 79%

The table below shows the accuracy of each method used above :

TABLE I. THE ACCURACY OF EACH METHOD USED T2

Method	**Accuracy**
(S1): Bayes with Laplace Smoothing	69%
(S2): Bayes with Laplace Smoothing and only relevant words	62%
(S3): Bayes with Laplace Smoothing and taking into consideration collocations when selecting relevant words	79%

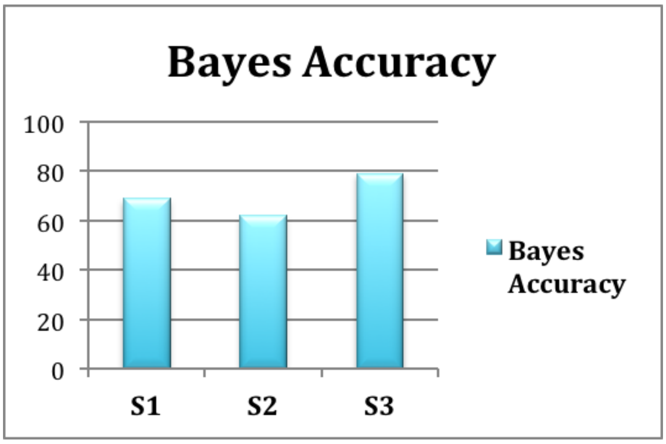


Figure 2. The accuracy of each Bayesien solution

As we can notice the third solution is the better one because it take into consideration the specificities of the Moroccan sociolect language.

B. *Support Vector Machine (SVM)*

The second classifier that we have applied to the corpus of the Moroccan sociolect language that we have built is the SVM classifier.

The Support Vector Machine (SVM) [13] was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is that SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers.

To apply the SVM classifier, we used WEKA software. WEKA (Waikato Environment for Knowledge Analysis) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Before using WEKA software, we defined a list of empty words or stop words that are specific to the Moroccan sociolect language, to eliminate them later from the corpus. And the next step is to represent the corpus in the form of vectors or word bag to train it later.

The SVM classifier application gave the following accuracy:

Accuracy (SVM) = 86%

C. *K - Nearest Neighbors (KNN)*

The third classifier that we have applied to the corpus of the Moroccan sociolect language is the KNN (k - Nearest Neighbors) classifier.

The KNN algorithm [14] is a method for classifying objects based on closest training examples in the feature space. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.

The goal of the KNN is to classify target points (whose class is unknown) according to their distances from points constituting a learning sample (i.e. whose class is known in advance).

The principle of the KNN is as follows: every unknown class data is compared to all stored data, and we choose for the new data the majority class among its K nearest neighbors.

The equivalent of the KNN classifier in WEKA is the IBK (Instance Based Learner) algorithm, and WEKA offers different types of distance to apply the KNN:

Hereinafter the accuracy result of the KNN according to a given distance:

- Euclidean distance

Accuracy (KNN/ Euclidean) = 61% (K1)

- Manhattan distance

Accuracy (KNN/ Manhattan) = 61% (K2)

- Chebyshev distance

Accuracy (KNN/ Chebyshev) = 53% (K3)

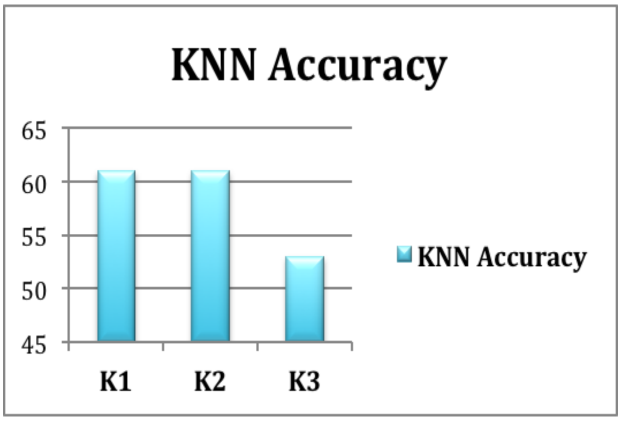


Figure 3. The accuracy of each KNN distance solution

As the diagram shows, when applying the KNN classifier, it is better to use either Euclidean or Manhattan distance to get a better accuracy.

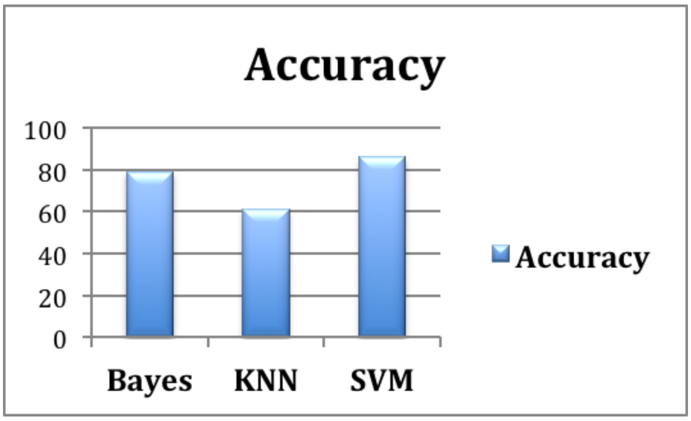


Figure 4. Comparaison of the accuracy of Bayes, KNN and SVM classifier

As we can conclude, the SVM classifier gave the best classification result.

VI. CONCLUSION AND PERSPECTIVES

In this paper we discussed this new trend of communication, that Moroccan Internet users use on social Media, that we called code-mixing which is a wide spread phenomena. Thus, in this context we have studied and applied several linguistic classifiers namely: Naive Bayes, SVM and KNN, and the comparison of these three classifiers showed that the SVM gives a better classification result.

Also, we can conclude that in order to get better results when applying the Bayes classifier we recommend using the Laplace Smoothing method and taking into consideration the collocations, specific to the Moroccan sociolect language, when defining the relevant corpus words.

And as a next step to this work we will work on the DELSOM dictionary to enrich it and add a section of polarity, also we will try to use deep learning to have a better classification results.

REFERENCES

1. Al Heeti, Niemahamad & Al Abdely, Ammar. (2016). TYPES AND FUNCTIONS OF CODE-SWITCHING IN THE ENGLISH LANGUAGE USED BY IRAQI DOCTORS IN FORMAL SETTINGS. 1
2. W. Abdelouahad Mabrour, Les langues au/du Maroc : une présentation sociolinguistique *Le français à l'université* , 21-01 | 2016 , on-line the : 14 March 2016, (09/02/2019, 13 :24)
3. Abdelouahad Mabrour, Les langues au/du Maroc : une présentation sociolinguistique *Le français à l'université* , 21-01 | 2016 , on-line the : 14 March 2016, (09/02/2019, 13 :24)
4. Sadiqi, Fatima. (1991). The spread of English in Morocco. International Journal of The Sociology of Language. 1991. 99-114. 10.1515/ijsl.1991.87.99.
5. Abderrahman Zouhir. Selected Proceedings of the 43rd Annual Conference on African Linguistics, ed. Olanike Ola Orie and Karen W. Sanders, 271-277. Somerville, MA: Cas-cadilla Proceedings Project
6. F.Mansouri, S.Abdelalim, El A.Ikram: A Modeling Framework for the Moroccan sociolect recognition used on the social media.BDCA 2017: 34:1-34:5, Proceedings of the 2nd international Conference on Big Data, Cloud and Applications Article No. 34
7. The annual Report The National Agency for the Regulation of Telecommunications ANRT-2015, https://www.anrt.ma/lagence/actualites/rapport-annuel-2015, (10/06/2017)
8. Alexa Ranking : statistics on the most visited websites in Morocco, http://www.alexa.com/topsites/countries/MA, (01/05/2017, 16:39)
9. Facepager : Data extraction software : https://github.com/strohne/Facepager, (05/08/2017, 15:42)
10. F.Mansouri, S.Abdelalim, El A.Ikram: A Modeling Framework for the Moroccan sociolect recognition used on the social media.BDCA 2017: 34:1-34:5, Proceedings of the 2nd international Conference on Big Data, Cloud and Applications Article No. 34.
11. F.Mansouri, S.Abdelalim, Y.Tabii. (2018) Modeling and Development of the Linguistic Knowledge Base DELSOM. In: Tabii Y., Lazaar M., Al Achhab M., Enneya N. (eds) Big Data, Cloud and Applications. BDCA 2018. Communications in Computer and Information Science, vol 872. Springer,Cham
12. A.FELDSTEIN, Overflow, Underflow, and Severe Loss of Significance in Floating-Point Addition and Subtraction, IMA Journal of Numerical Analysis (1986) 6, 241-251
13. K. SRIVASTAVA, L.BHAMBHU, « DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE », Journal of Theoretical and Applied Information Technology, 2005 - 2009 JATIT
14. S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background”.