

Sentiment Analysis of COVID-19 Tweets: Impact of Pre-processing Step

Rami Ayadi^{1,3}, Osama R. Shahin¹, Osama Ghorbel¹, Rayan Alanazi¹ and Anouar Saidi^{2,4}

ravyadi@ju.edu.sa; orshahin@ju.edu.sa; oaghorbel@ju.edu.sa; rmalanazi@ju.edu.sa; aaalsaidi@ju.edu.sa

¹Department of Computer Sciences, College of Science and Arts of Gurayyat, Jouf University, Saudi Arabia

²Department of Mathematics, College of Science and Arts of Gurayyat, Jouf University, Saudi Arabia

³University of Gabes, Isim mednine, Tunisia

⁴University of Monastir, FSM, Tunisia

Summary

Internet users are increasingly invited to express their opinions on various subjects in social networks, e-commerce sites, news sites, forums, etc. Much of this information, which describes feelings, becomes the subject of study in several areas of research such as: "Sensing opinions and analyzing feelings". It is the process of identifying the polarity of the feelings held in the opinions found in the interactions of Internet users on the web and classifying them as positive, negative, or neutral. In this article, we suggest the implementation of a sentiment analysis tool that has the role of detecting the polarity of opinions from people about COVID-19 extracted from social media (tweeter) in the Arabic language and to know the impact of the pre-processing phase on the opinions classification. The results show gaps in this area of research, first of all, the lack of resources when collecting data. Second, Arabic language is more complexes in pre-processing step, especially the dialects in the pre-treatment phase. But ultimately the results obtained are promising.

Key words:

C.ovid-19, Sentiment Analysis, Covid-19 Tweets, SVM, KNN, pre-processing step, Arabic Language.

1. Introduction

The emergence of more and more disasters and epidemics around the world is prompting people to write more and more about these abrupt events. Some of these disasters have deprived human beings of their basic rights such as the right to meet and discuss with others face to face, which pushes people to protest remotely behind their protest screens on social networks and to express their opinions and feelings about a given event.

Since January 2020, an epidemic of COVID-19 Coronavirus (ex 2019-nCoV) has spread from China. Then it spread around the world to affect 210 countries and territories and even affect international means of transport. But the number of people affected by COVID-19 is increasing rapidly and exceeds 100 million cases, including more than 2 million deaths, making it an epidemic for this virus.

Social networks amuse a major role in conveying people's different opinions about this epidemic, as people express their feelings about how the virus is being transmitted, about the different preventive measures being taken by governments, about how the different ministries of health disseminate the right and true information.

This flow of information contained in the writings of people in social networks has encouraged many researchers to be interested in the analysis of sentiment. The core of this work is to an analysis of sentiment in Arabic texts found in social media.

Among the many social networks available on the net, Twitter is the best known and used in the Arabic world [21]. People contact their relatives, friends, government services, also to write comments and express their feelings on various subjects. By analyzing the feelings, we will give the texts written remotely (tweets) a class. This class expresses people's feelings which can be neutral, negative, or positive.

Sentiment analysis (SA) is performed using supervised machine learning algorithms. These algorithms can predict the class to which new unclassified documents such as tweets may belong after a learning phase [19].

A little recent, the field of Sentiment Analysis attracts the attention of some research [20]. It can grant a great profit to the company what encourages them to finance this field. The analysis of feelings is one of the problems appeared with the automatic processing of languages, the objective of this research evidence is the analysis of the opinions of Internet users and to discover the polarity (positive / negative) of a given subject

Therefore, in this work we have chosen to focus on people's opinions on the new epidemic that has appeared (covid-19) and more specifically on the comments written on tweeter. The objective is to know the impact or the effect of the pre-processing phase on the classification of the opinions of the Net surfers concerning covid-19 in positive or negative.

The majority of the work done in this area has been on European, English and Asian languages. To date, few works have been published for morphologically complex languages, specifically the Arabic language.

In this research, we will present the majority of the problems encountered by researchers during the polarity detection phases and who work on corpuses in the Arabic language to propose a set of steps to follow to overcome these difficulties.

First, we will present some previous work on the analysis of feelings concerning industrial products. then the phase of "data collection" from tweeter, then we will simplify the different steps carried out during the pre-processing phase, to then perform the classification and detect the polarity of people's opinions concerning Covid-19. Classification is ensured by applying the three most used algorithms to know the SVM, KNN, NB and finally discuss the impact of the pre-treatment steps on the results found in the classification.

2. Related Works

As in other languages, many experiences on sentiment analysis in Arabic that exist. they use a variety of Learning Machine. The most used for Arab sentiment analysis are SVM and Naive Bayes.

The detection of subjectivity and feelings was the goal of early work in the task of sentiment analysis in Standard Arabic [2].

In [3] and [4], we find the first essay for the construction of a corpus designed for subjective analysis and sentiment analysis then they introduced an SA system for social media. A wide range of features is experienced in their study. Moreover, the morphological richness of Arabic is studied in [5] with different possibilities to manage it.

In [6], the authors conducted a comparative study on the performance of SVM to an RNN-based model. This study has the objective of building a system of analysis of feelings based on the aspect Sentiment analysis based on the feeling linked to the characteristics or service of a product by identifying the aspects. The model is tested on data containing hotel reviews written in Arabic.

In [7], the results in their approach showed that SVM (with an accuracy of 95%) exceeds the RNN model (with an accuracy of 87%). in this study, the authors combined lexical, syntactic, semantic and morphological characteristics of the Arabic language for sentiment analysis.

In [8], the authors built a sentiment analysis system in Arabic. This system is tested on data collected and tagged manually from tweets. The authors used their own lexicon. The precision achieved is of the order of 87%.

In [9], the authors targeted social media where they worked with dialect variations in social media by putting together their own lexicon like sentimental slang words and lexicon idioms (SSWIL). in this work, an accuracy of

87% is found using the lexicon combined with an SVM classifier.

In [10], the author experimented on 2,000 manually collected tweets different sentiment analysis approaches. The best result found was 87% accuracy with the SVM classifier.

In [11], the authors presented a system of sentiment analysis on tweets dealing with health services. They experimented with the SVM, Naive Bayes and CNN classifiers on the collected dataset. The best result is obtained with the SVM classifier (an accuracy of 91%).

3. Data Collection

We collected corpus from twitter. We used the RTweet package in R and we used as keywords COVID19, COVID-19, كورونا, جانحة كورونا, كوفيد-19, لأجل سلامتک ابقی فی المنزل.

Then a preliminary filtration step is carried out on the reweets and the responses during the collection phase such as removing duplication of tweets, white spaces, punctuation, and stop words. The corpus contains 1,500 documents, 20,000 sentences and 45,000 words.

4. Corpus Labeling

Labeling or annotating feelings is a problem that needs human effort. To guarantee the success of this step, we used two evaluators to annotate the corpus, a first expert in health, and a second specialist in the Arabic language (teacher). We do that to guarantee the process of their classification. If we have a conflict in a document, we used a third non-specialist assessor to validate the choices. In the Final, we have a corpus that contains 1,500 documents annotated like this, 725 positive and 725 negative.



5. Major Problem

5.1 Emoji

We have noticed that users of social networks introduce smileys in their comments because the smileys save time in writing and it better represents the emotional state of the user as expressing joy can be reduced to the use of a smiley, also anger and a lot of other emotion. For this, we have introduced a conversion step whose objective is to convert the smileys into text) as shown in the following table 1.

Table 1. Convert emoji to text

smileys	Conversion
---------	------------

	Grin ابتسامة
	Cool ممتاز

5.2 Extended words

We also noticed that users of social networks repeat the characters in a word. After an analysis, we found that this repetition aims to express either an assertion or an emphasis. For this we have developed a correction algorithm to eliminate the redundancy of the characters but moreover we have tried to keep the meaning of the words without modification. Example:

رائع (good)	رائعـــــــــــــــــــــــــــــــــ (goooooood)
-------------	---

5.3 Use of dialects

One of the major problems of the Arabic language is the number of dialects existing even in the same country. Users of social networks are more relaxed if they write in their own dialect to express an opinion. The use of these dialects complicates the task of analysing feelings for the Arabic language.

5.4 Bi-lingual comments

We can also find comments that contain characters or words in a foreign language than the Arabic language like English or French. Example

من احسن الشاشات في هذا الموسم it's good
A good هاتف

5.5 Off topic comments

Users write several comments on a lot of topic in and sometimes they write on the same topic in several places but these comments are off topic or it does not express an opinion on the topic in question

To do this, we have introduced a step in the pre-treatment process which will be able to eliminate these unrepresentative noises

6. Classification

The classification process consists of analyzing new data and then assigning it to predefined classes based on their characteristics [14] [15].

This part will be devoted to present the details of the experiments carried out. The first step is the derivation

phase in which we will remove all the prefixes and suffixes from a word to produce the root

Using the Arabic Stemmer makes detecting the polarity of these words a difficult task, because in Arabic from the same root, we can generate various word forms that do not have a similar meaning. For example, the stemming of the two words "رائع" (wonderful) and "مروع" (terrible) gives the word "روع" (horror). We can see that the polarity of "رائع" (wonderful) is reversed by the stemming [13]. However, when using light stemming the task became to strip a small set of prefixes and/or suffixes while retaining the meaning of the words. For example, the word "المسافرون" (travelers) change to "مسافر" (traveler) not the root "سفر" (Travel).

It is important to mention the existence in a third algorithm similar to the Arabic stemmer algorithm called 'Arabic Khoja stemmer', it is developed by Shereen Khoja [16] whose role is to remove the longest suffixes and the prefixes. It then compares the remaining word with verbal and nominal patterns for the extraction of the root. The stemmer uses several linguistic data files such as a list of all diacritical characters, punctuation characters, specific articles, and 168 stop words.

We tested the impact of three types of stemmers, Arabic Stemmer, Arabic Light Stemmer, and Khoja stemmer.

7. System Architecture

Our prototype has a modular architecture. Its main tasks are as follows: collecting comments on the Internet from tweeter, then pre-processing the data, then the stemming phase where the words are reduced to their roots, and finally the detection of the polarity of opinions according to the SVM, Naïve Bayes (NB), K-nearest neighbour (KNN) classifier either positive or negative.

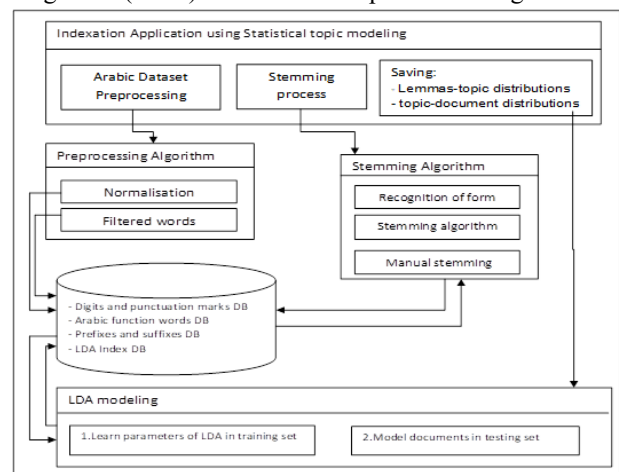


Fig. 1 General architecture of sentiment analysis system

8. Result and Discussion

We did our tests with several types of classification algorithms. The Vector Machine Support (SVM), Naive Bayes (NB), K-nearest neighbour (KNN) are cited as an example. And these are applied to different combinations of pre-processed data. The classifier performance testing phase applied to the corpus is done using the free weka tool it is a popular suite of machine learning software. Written in Java, developed at the University of Waikato, New Zealand [17].

To perform the performance tests, there are several techniques; we chose two among them which are: Cross-validation and split percentage

8.1 Cross-Validation

Cross-validation is a very popular alternative for managing the sparse nature of data [18]. This involves dividing the data set into K groups drawn at random. These groups work well as a test set. We can then calculate and average the test error for each group, which is the estimator of the cross-validation test error. In this study, the number K is equal to 10, that is to say, the learning set is cut into ten groups.

We will learn the algorithm ten times over nine parts then we will evaluate the model on the remaining teeth. The 10 assessments are then joined.

The tables and figures below present the results obtained concerning the different classifiers in terms of precision and recall with the reliability estimation method "cross-validation or in English cross-validation

Table 1. Classification results by the 'cross validation' assessment method in terms of accuracy

<i>Corpus / Classifier</i>	<i>KNN</i>	<i>NB</i>	<i>SVM</i>
Standard Corpus	0.722	0.81	0.89
light stemmer	0.756	0.84	0.92
khoja stemmer	0.75	0.84	0.91
normalization	0.69	0.88	0.88
normalization+khoja stemmer	0.68	0.873	0.931
normalization+light stemmer	0.66	0.875	0.929

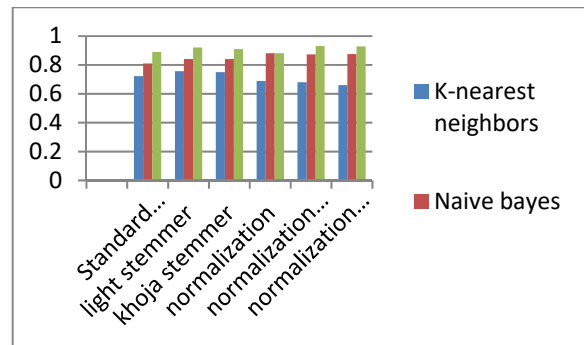


Fig 1. Classification results by the 'cross-validation' assessment method in terms of accuracy

Table 2. Classification results by cross validation evaluation method in terms of recall

<i>Corpus / Classifier</i>	<i>KNN</i>	<i>NB</i>	<i>SVM</i>
Standard Corpus	0.713	0.82	0.901
light stemmer	0.72	0.87	0.922
khoja stemmer	0.726	0.877	0.92
normalization	0.621	0.883	0.897
normalization+khoja stemmer	0.599	0.891	0.921
normalization+light stemmer	0.6	0.9	0.922

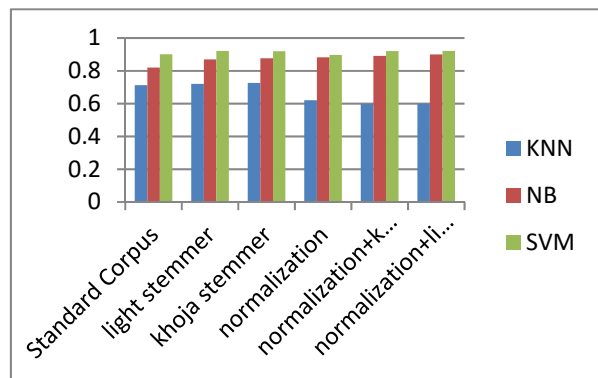


Fig 2. Classification results by cross-validation evaluation method in terms of recall

From Figures 2 and 3, we conclude that there is a classifier that has given poorer performance than the other two; it is the k-nearest neighbors (KNN) classifier. Unlike the results obtained by the Naive Bayes (NB) and the Support Vector Machines (SVM) which are more or less close. Except that the SVM is the most efficient with the different types of data combinations, we achieved 0.92 of precision and 0.922 of recall with the "cross-validation" reliability estimation method.

8.2 Percentage Split

We have randomly divided the corpus into two separate data sets. The first set is called the training set (which the system can extract knowledge). The test set is the second set, is the data to be categorized.

In this study, of the set of data training is equal to 80% and for the test we spent the rest 20%. The figures which follow present the results obtained for the various classifiers and precision and recall are used.

Table 3. Results of classification in terms of accuracy

Corpus / Classifier	KNN	NB	SVM
Standard Corpus	0.823	0.84	0.966
light stemmer	0.819	0.901	0.966
Khoja stemmer	0.818	0.901	0.968
normalization	0.798	0.942	0.919
normalization+khoja stemmer	0.821	0.966	0.95
normalization+light stemmer	0.822	0.966	0.951

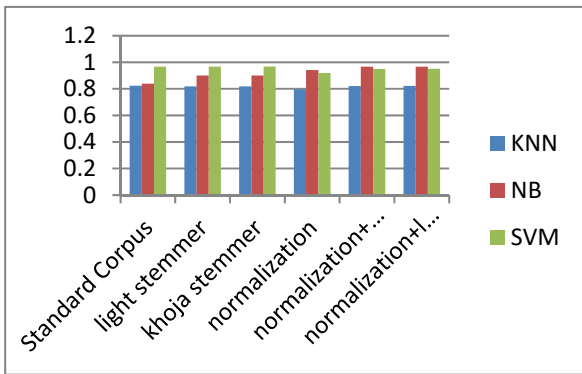


Fig 3. Results of classification in terms of accuracy

Table 4. Classification results in terms of recall

Corpus / Classifier	KNN	NB	SVM
Standard Corpus	0.796	0.835	0.96
light stemmer	0.755	0.899	0.961
khoja stemmer	0.755	0.899	0.962
normalization	0.734	0.937	0.917
normalization+khoja stemmer	0.677	0.956	0.917
normalization+light stemmer	0.678	0.955	0.929

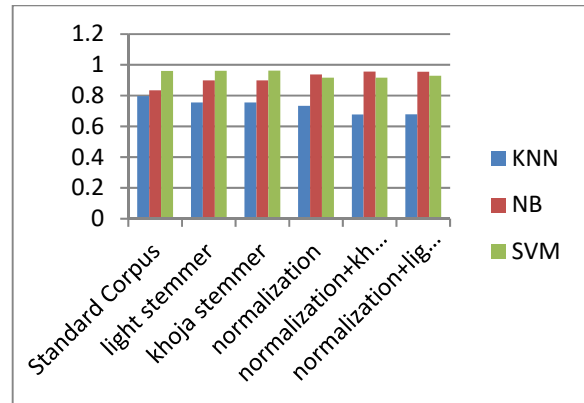


Fig 4. Classification results in terms of recall

Similarly concerning the KNN classifier when changing the evaluation method "split percentage", it is found that it did not manage to give close performance than the other two classifiers. This time the SVM won the challenge that on the raw corpus, the corpus with the light stemmer and the corpus with the Khoja stemmer, on the other hand, the NB is more effective with the combinations of data corpus more normalization, corpus more normalization more Khoja stemmer and corpus more normalization more light stemmer. We reached 0.948 of accuracy and 0.961 of recall.

8. Conclusion

The application of three classification algorithms Naïve Bayes (NB), Support Vector Machines (SVM), and K-nearest neighbors (KPPV) on a corpus in the Arabic language showed the superiority of the first two classifiers compared to the third. The results obtained show the impact of the pretreatment phase and the application of different stemming techniques compared to a set of data (product reviews) in various Arabic dialects. The assessment is made by the two techniques cross-validation and percentage split. The best details achieved are in the following cases:

- The use of Naïve Bayes with a standardized corpus plus the application of Khoja Stemmer or Light Stemmer to give 0.948 of accuracy and 0.961 of recall.
- The use of Support Vector Machines either with the raw corpus or with the corpus plus the application of Khoja Stemmer or Light Stemmer to also give 0.948 of accuracy and 0.961 of recall.

References

- [1] Dubey, A. D. (2020). Twitter Sentiment Analysis during COVID19 Outbreak. Available at SSRN 3572023..
- [2] M. Abdul-Mageed, M. Diab, M. Korayem .Subjectivity and sentiment analysis of modern standard Arabic. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Association for Computational Linguistics, Portland, Oregon, USA (2011), pp. 587-591.
- [3] M. Abdul-Mageed, M. Diab. AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. Proceedings of the eighth international conference on language resources and evaluation (LREC 2012), European Language Resources Association (ELRA), Istanbul, Turkey (2012), pp. 3907-3914
- [4] M. Abdul-Mageed, M. Diab, S. Kbler. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28 (1) (2014), pp. 20-37.
- [5] M. Abdul-Mageed. Modeling arabic subjectivity and sentiment in lexical space. *Information Processing & Management*, 56 (2) (2019), pp. 291-307.
- [6] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, B. Gupta. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels reviews *Journal of Computational Science*, 27 (2018), pp. 386-393.
- [7] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, G. Eryigit. SemEval-2016 task 5: Aspect based sentiment analysis; Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California (2016), pp. 19-30.
- [8] M. Al-Ayyoub, S.B. Essa, I. Alsmadi. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining*, 2 (2) (2015), pp. 101-114.
- [9] T.H. Soliman, M.A. Elmasry, A. Hedar, M.M. Doss. Sentiment analysis of arabic slang comments on facebook. *International Journal of Computers & Technology*, 12 (5) (2014), pp. 3470-3478.
- [10] N.A. Abdulla, N.A. Ahmed, M.A. Shehab, M. Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (aeect) (2013), pp. 1-6
- [11] A.M. Alayba, V. Palade, M. England, R. Iqbal. Arabic language sentiment analysis on health services. 2017 1st international workshop on arabic script analysis and recognition (asar) (2017), pp. 114-118
- [12] Mountassir, A., Benbrahim, H., & Berrada, I. (2012). A cross-study of Sentiment Classification on Arabic corpora. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 259-272). Springer, London.
- [13] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López et J. M. Perea-Ortega, (2011)«Bilingual Experiments with an Arabic-English Corpus for Opinion Mining,» Proceedings of Recent Advances in Natural Language Processing, p. 740–745.
- [14] R. Ayadi, M. Maraoui et M. Zrigui,(2009) «Intertextual distance for Arabic texts classification,» ICITST, pp. 1-6.
- [15] Zrigui, R. Ayadi, M. Mars et M. Maraoui, (2012) «Arabic Text Classification Framework Based on Latent Dirichlet Allocation,» *Journal of Computing and Information Technology - CIT* 20, vol. 2, p. 125–140.
- [16] S. Khoja,(2002)«Shereen Khoja - Research,» [En ligne]. Available: <http://zeus.cs.pacificu.edu/shereen/research.htm>.
- [17] Machine Learning Group at the University of Waikato(2013), «Weka 3: Data Mining Software in Java,». [En ligne]. Available: <http://www.cs.waikato.ac.nz/ml/weka/index.html>.
- [18] J. Chiquet, (2009) «Validation croisée pour le choix de paramètre de méthodes,» *Module MPR – option modélisation*.
- [19] [15] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. Javier González-Castaño, “Unsupervised method for sentiment analysis in online texts,” *Expert Systems with Applications*, vol. 58, pp. 57–75, 2016
- [20] GOEL, Ankur, GAUTAM, Jyoti, et KUMAR, Sitesh. Real time sentiment analysis of tweets using Naive Bayes. In : 2016 2nd International Conference on Next Generation Computing Technologies (NGCT). IEEE, 2016. p. 257-261.
- [21] AHMAD, Munir, AFAB, Shabib, et ALI, Iftikhar. Sentiment analysis of tweets using svm. *Int. J. Comput. Appl*, 2017, vol. 177, no 5, p. 25-29.