The 2022 International Conference on Digital Technologies Applications (ICDTA'22)

Leveraging the Automated Machine Learning for Arabic Opinion Mining: A preliminary study on AutoML tools and comparison to human performance

Moncef Garouani[1,2,3] , Kasun zaysa[1]

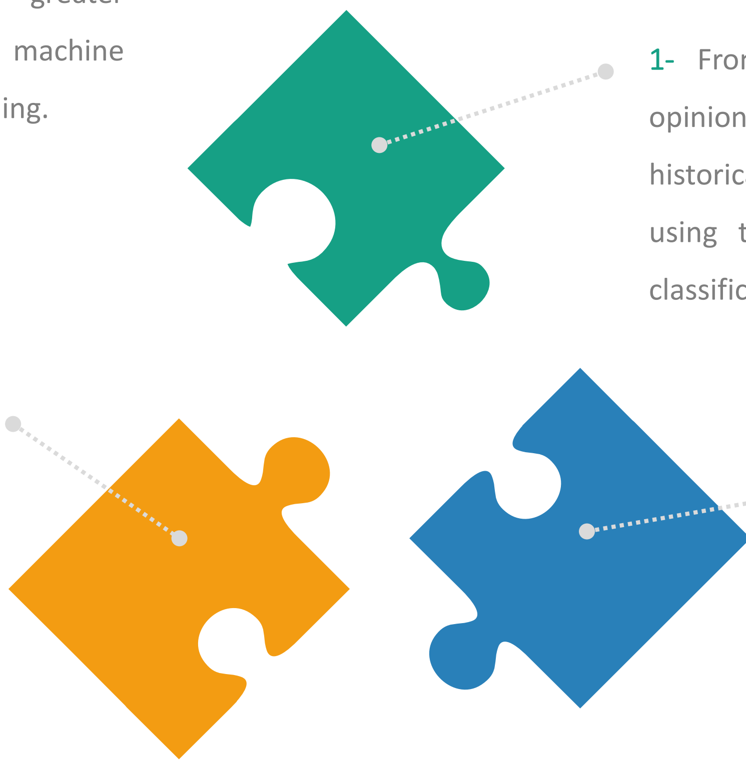
1. LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France
2. CCPS Laboratory, ENSAM, University of Hassan II, Casablanca, Morocco
3. Informatics Institute of Technology, University of Westminster Colombo, Sri Lanka

Motivation

With the advent of the web 2.0 and the explosion of data sources such as review platforms, blogs and microblogs, there has been a need to analyze millions of posts, tweets or reviews in order to find out what internet users think.

Motivation

Increasing data availability and greater computing capacity have enabled machine learning (ML) to address opining mining.



1- From a machine learning perspective, opinion mining is a technique that uses historical data to create predictive models using textual data to make predictive or classification decisions.

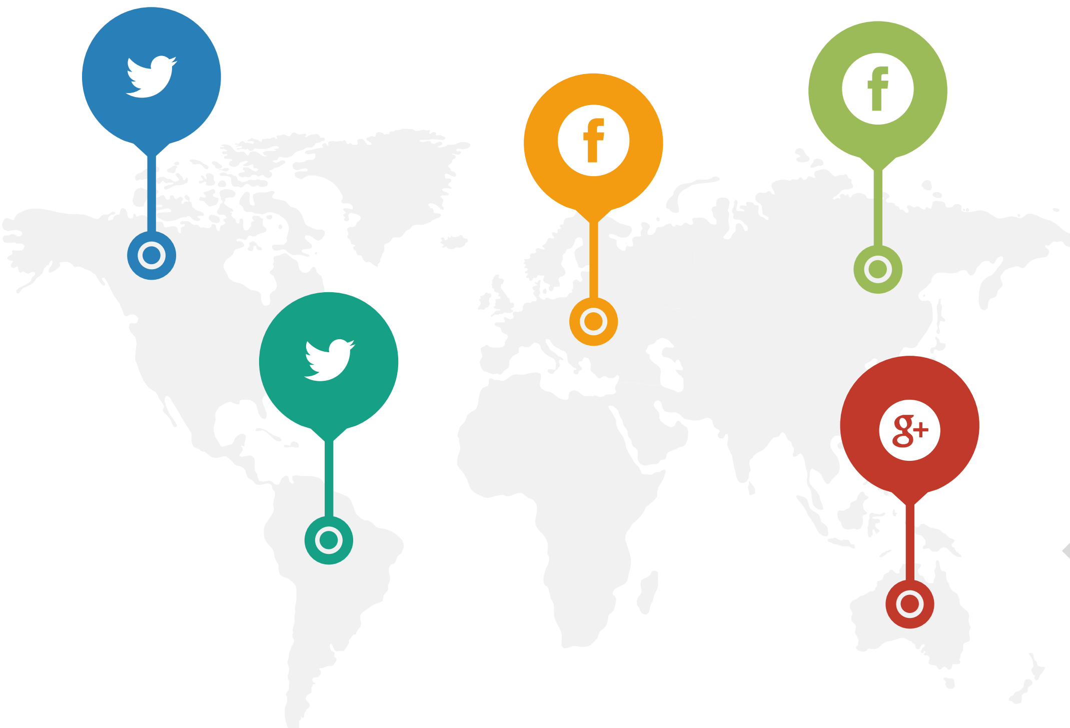
2- There is no ML algorithm that would perform well across all types of textual data.

3- Determining the more adequat ML method or algorithm for the problem at hand is a complex task that requires expert ML knowledge

PLAN

- Introduction
- Opinion mining
- Automated machine learning
- Methodology
- Experiments and Results
- Conclusion & Perspectives

Introduction



Social media

Facebook, Twitter, Instagram, LinkedIn, these social platforms are now part of everyday life. The data aspect of these social media, such as Twitter messages, generates a rich wealth of data about who is involved in communication.

120K **530K** **319K**

This data plays an important role in decision making for many people and organizations.

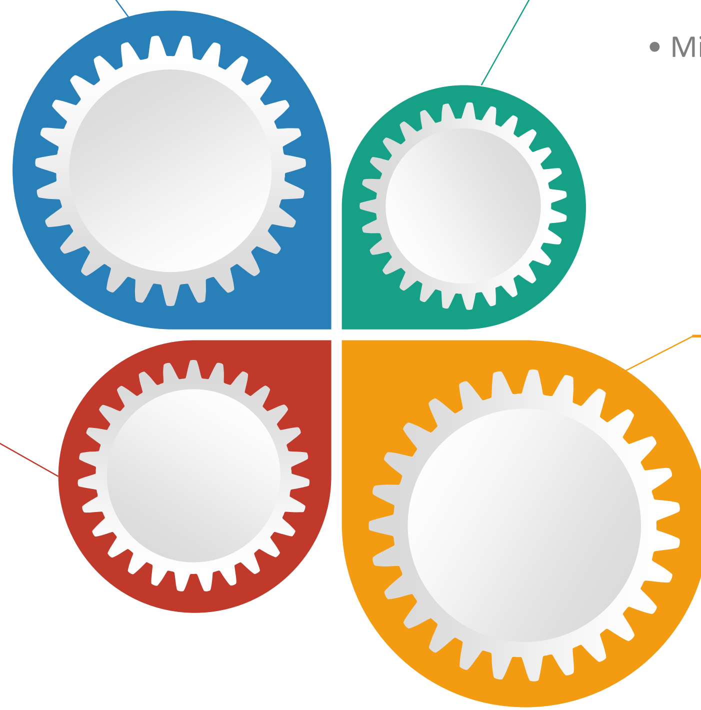
Opinion mining

Opinion mining

Refers to technologies for the automatic analysis of speech, written or spoken, in order to extract subjective information such as judgments, evaluations or emotions.

Data Sources

- Review sites
- Blogs
- Micro-blogs: Twitter, Facebook…



Approaches

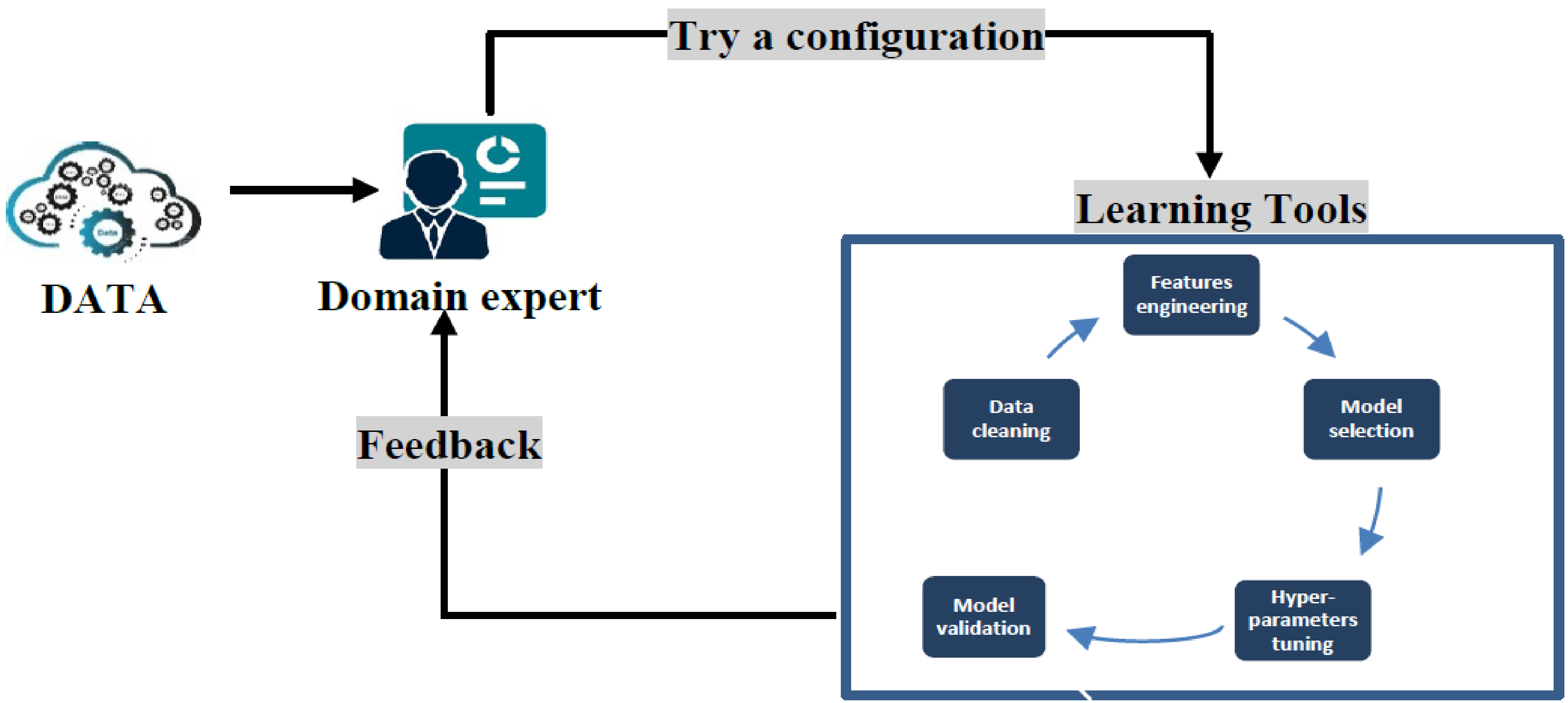
- Machine Learning Approach
- Lexicon-based / dictionary rule-based methods (Semantic orientation)

Application areas

- Politics / political science
- Commercial
- Sociology
- Finance

Opinion mining

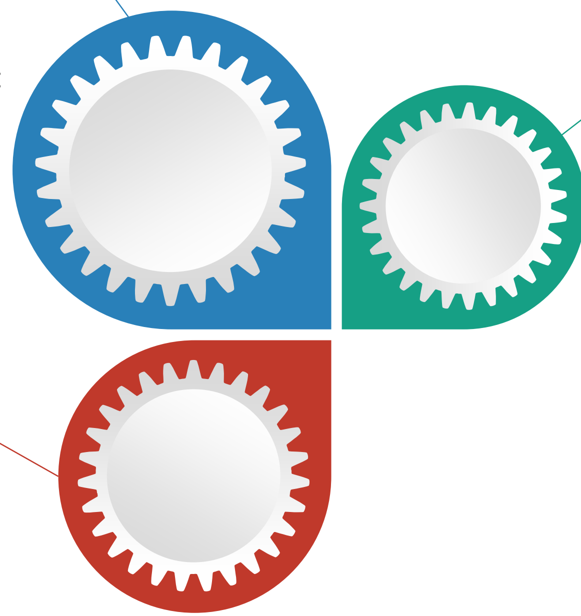
Algorithms selection problem



Automated machine learning

AutoML

AutoML aims to find or identify the optimal set of preprocessing techniques, ML algorithms, and HPs to maximize a performance criterion on the data without being specialized in the problem domain where the data comes from.



Application areas

- Educational data analysis
- Health care applications
- Manufacturing industry

AutoML tools

- **[1]** Auto-Sklearn
- **[2]** TPOT
- **[3]** AutoWeka
- **[4]**AMLBID

[1] M. Feurer et al. Eficient and Robust Automated Machine Learning". In: Proceedings of the 28th International Conference on Neural Information Processing Systems
[2] R. S. Olson and J. H. Moore. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning". doi: 10.1007/978-3-030-05318-5.
[3] L. Kottho et al. Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA. doi: 10.1007/978-3-030-05318-5.
[4] Garouani, M et al. (2022). AMLBID: An auto-explained Automated Machine Learning tool for Big Industrial Data. In SoftwareX . doi: 10.1016/j.softx.2021.100919

V- Methodology

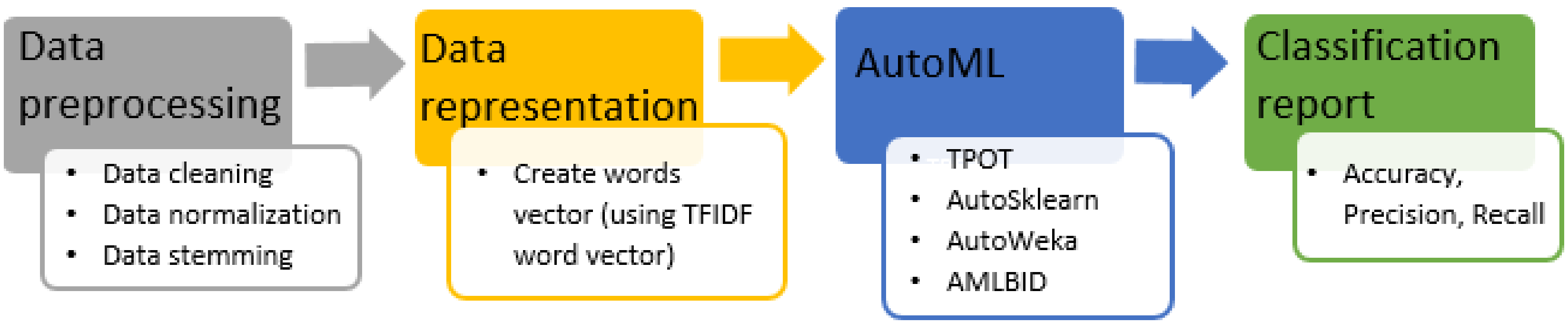


Fig. 1. Evaluation Flow.

V- Methodology

1. Case study and raw data

Dataset	Number of Instances	Number of classes	Arabic
D1	13350	4	Moroccan
D2	49864	2	Algerian
D3	900	2	Jordanian
D4	9901	2	Moroccan
D5	17000	2	Tunisian
D6	510600	3	Egyptian
D7	4462	2	Arabic
D8	66666	3	Arabic
D9	56862	2	Arabic
D10	11751	2	Egyptian

Table 1: Datasets description

V- Methodology

2. Data preparation and representation

All datasets were prepared in such a way that only two columns remained text and target column. In order to do so, a pre-processing stage is done to minimize the effect of text informality on the classification. The pre-processing stage includes **Emojis removal**, **repeated letters elimination**, Arabic characters **normalization** and finally the **stemming**. For the features representation, we used the **TFIDF** representation.

Analysis evaluation

Dataset	AutoML results				Human configuration results
	TPOT	Austo-Sklearn	Auto-Weka	AMLBID	
D1	85.73%	90%	87.07%	91.47%	**92.09%**
D2	**86.71%**	65.89%	50.21%	83.91%	86%
D3	81%	79.22%	75.07%	80.62%	**86.89%**
D4	73.11%	83.91%	80.49%	79.46%	**84.33%**
D5	86.02%	83.99%	69.21%	**89.71%**	78%
D6	**79.55%**	80.36%	68.67%	77%	79.05%
D7	88.51%	83.72%	65%	**88.80%**	87%
D8	79.30%	80.73%	70.45%	87.03%	**91%**
D9	53.09%	**78.16%**	62.51%	73.61%	77%
D10	80.74%	82%	65.23%	**86.37%**	78%

Table 2. Performance results of each evaluated method.

Analysis evaluation

Approche Machine Learning

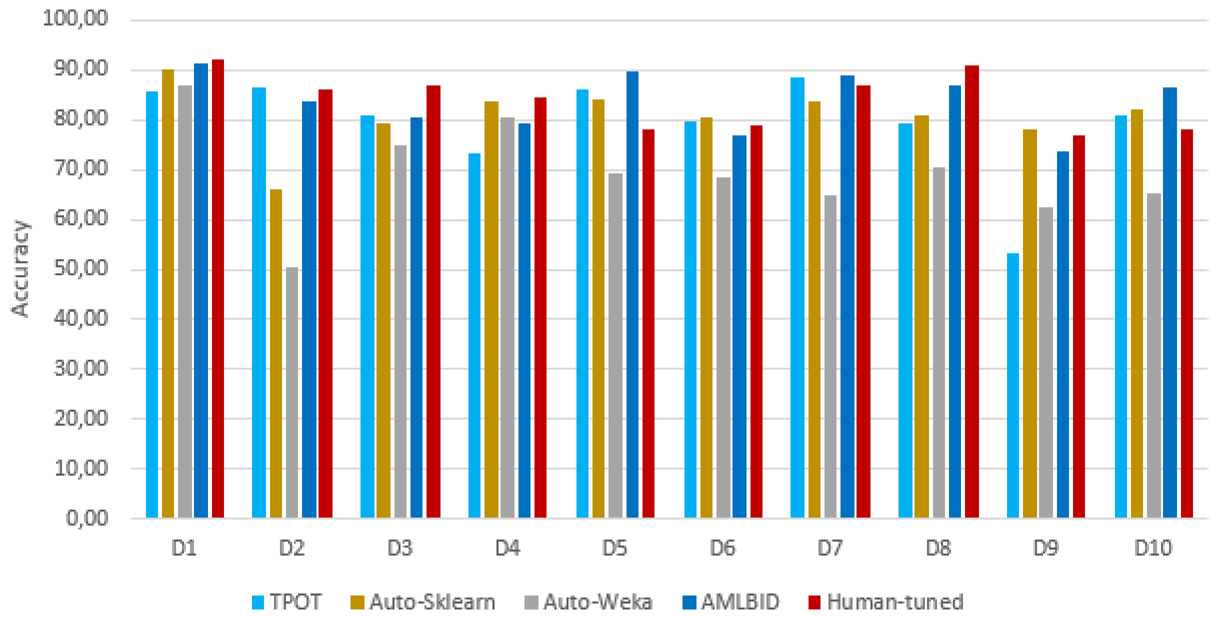


Figure.2. Comparative results of the effectiveness of AutoML over default classic ML configurations and domain expert (Human) configurations.

Conclusion

This study has investigated the effectiveness of AutoML techniques on the Arabic opinion mining.

We explored the benefits of using AutoML in the OM field.

we have empirically studied to what extent the results of AutoML differ from the general approach of OM.

We used Auto-WEKA, TPOT, AutoSklearn and AMLBID as AutoML tools on **10** benchmarked datasets.

We performed several scenarios using several parameters:
N-grams, Stopwords removal, and TF-IDF

Experimental results show that the AutoML technology can be considered as a powerful approach to support the ML algorithm selection problem in OM.

Perspectives

The next planned steps include:

1. Compare the performance of more AutoML tools in deferent opinion mining cases to assess the consistency of the AutoML .

2. Carry out a pre-processing and features importance study to determine what are the attributes of the Arabic text that more influence the performance of AutoML.

THANK YOU FOR YOUR ATTENTION

To your questions

UNIVERSITÉ DU LITTORAL CÔTE D'OPALE

ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS
UNIVERSITÉ HASSAN II DE CASABLANCA

The 2022 International Conference on Digital Technologies Applications (ICDTA'22)

Leveraging the Automated Machine Learning for Arabic Opinion Mining: A preliminary study on AutoML tools and comparison to human performance

Moncef Garouani[1,2,3] , Kasun zaysa[1]

1. LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France
2. CCPS Laboratory, ENSAM, University of Hassan II, Casablanca, Morocco
3. Informatics Institute of Technology, University of Westminster Colombo, Sri Lanka