Arabic Twitter Detection for Misogynistic Language

Ashraqat Torky
atork095@uottawa.ca

Mahmoud Helmy
mhelm081@uottawa.ca

Sarah Elmasry
selma083@uottawa.ca

Yomna Abdelsatar
yabde005@uOttawa.ca

***Abstract*—Social media platforms have given people the freedom to speak their minds. However, as the media evolves, hate speech also evolves. Misogynistic language is a form of hate speech. It is reflecting or exhibiting hatred, dislike, or mistrust of women. To provide a more safe and free of abusive environment, such language should be labeled so that proper actions can be taken. This project's goal is to auto detect and analyze Arabic tweets that contain misogynistic language. We used Arabic Levantine Twitter dataset for Misogynistic language (LeT-Mi) released in 2021. By trying out various Natural Language Processing techniques, we reached a champion combination of preprocessing, feature extraction and modeling to achieve optimal performance results.**

I. Introduction

In many countries, most of the victims of online hate speech are members of minority groups. Women belonging to these groups are disproportionately targeted. Online gender-based violence can have significant psychological, social, and economic impacts. In addition to directly impacting the women who are present online, online gender-based violence could be predictive of violent crimes in the physical world. Most directly, it affects women's freedom of expression.

From the United Nations to local non-profit organizations, entities across different geographies have actively acknowledged the pervasiveness of online gender-based violence. Yet, major gaps remain, in research as well as in implementation of interventions to minimize its prevalence. As members of the tech community, we realized it is important to hold the tech industry responsible for creating platforms that are more conducive to women's participation. Furthermore, as being speakers of a complex language like Arabic, we determined to use our linguistic knowledge to support this goal.
Arabic is one of the most spoken languages in the world. It is morphologically rich with complex grammatical structure and intricate sentence structure. This nature, beside the lack of benchmark datasets form a challenge when it comes to Arabic NLP applications.

In this project we used Arabic Levantine Twitter dataset for Misogynistic language (LeT-Mi) [1] which is the first benchmark dataset for Arabic misogyny released in March 2021. It is annotated based on the misogynistic behavior either as neutral (misogynistic-free) or as one of seven misogyny categories: discredit, dominance, cursing/damning, sexual harassment, stereotyping and objectification, derailing, and threat of violence.

The goal is to auto detect misogynistic language in Arabic text, and classify it to one of previously mentioned 7 categories. Moreover, extract insights from them to further analyze the phenomenon.

II. Data Presentation

A. Data Description

The dataset consists of 5239 tweets annotated either as normal (misogynistic-free) or misogyny categories: discredit, dominance, cursing/damning, sexual harassment, stereotyping and objectification, derailing, and threat of violence. We notice

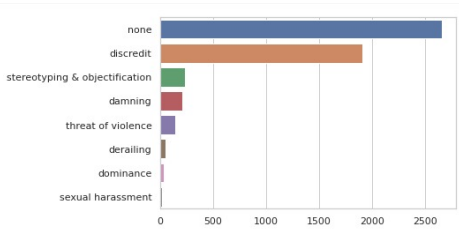


Fig. 1: Category Distribution

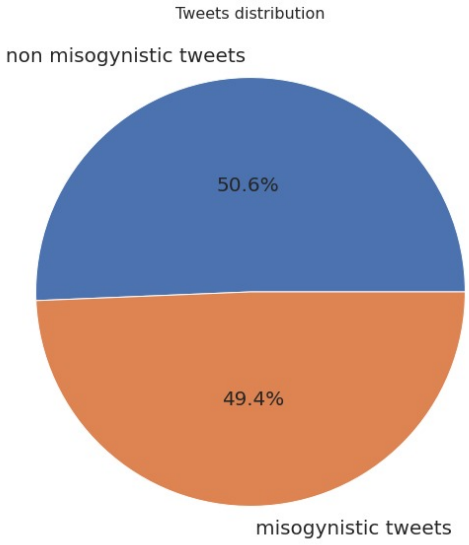


Fig. 2: binary classification

that the categories are unbalanced which is common problem at twitter data. we apply oversampling approach called SMOTE.

B. Data Preprocessing

Due to the uniqueness of the historical and cultural background of the Arabic language, its nature and structure are different from other languages such as English. For example, this language is written from right to left.Also Arabic includes 28 letters,three vowels and diacritics.
The result, we have several steps in the preprocessing and cleaning:

- twitter mentions like @username.
- links or websites
- Arabic and English punctuation
- retweet 'RT'
- repeated letters
- remove mobile emojis
- Tokenization

C. Data Visualization

1) Word Cloud: To visualize the dataset , we used word cloud to visualize the most frequent words in the data set.

Fig. 3: words frequent in the dataset

form the above figure we can see that's some of words are insults, this illustrate the pain and the discrimination that women see in different social media platform, some of the words in translation to English : idiot woman, woman who lost her sexual morals and many other words.

Fig. 4: words frequents in non misogynistic tweets

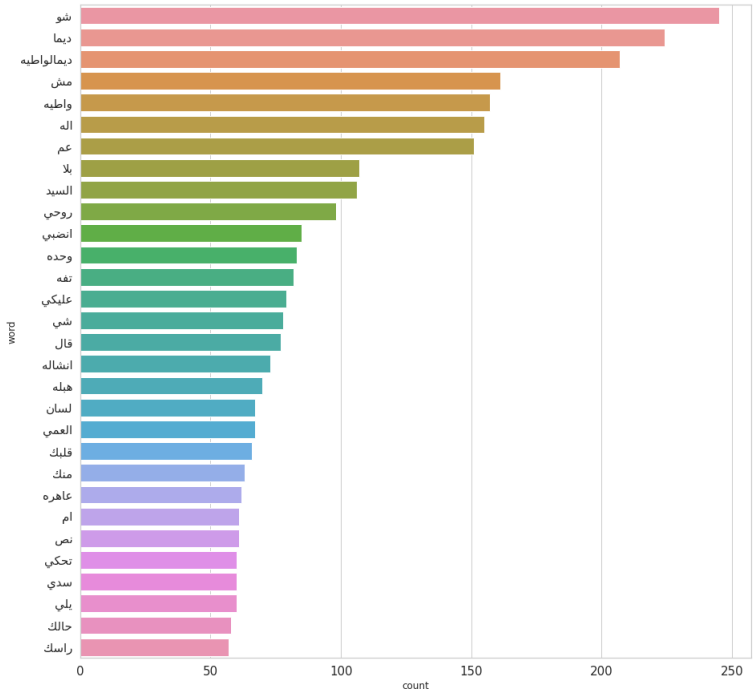


Fig. 5: bar-chart for all tweets

2) Bar-chart plots: To visualize more our dataset,we used the bar-chart plots. from the plot we see that the third most repeated words is an insults which means that woman has declasse morals, the 11th words is a mean words which means behave.

Conclusion:
we saw how the people treats women in the social media platform ,which appears in the most frequent or repeated words in the dataset and that appears as serious problem .

III. Classification

To predict which misogyny category the tweet belongs to, three different algorithms are used: Support Vector Machines (SVM), k-Nearest Neighbor, and Decision Tree using two different feature engineering techniques: Bag of Words and TF-IDF.

A. Evaluation

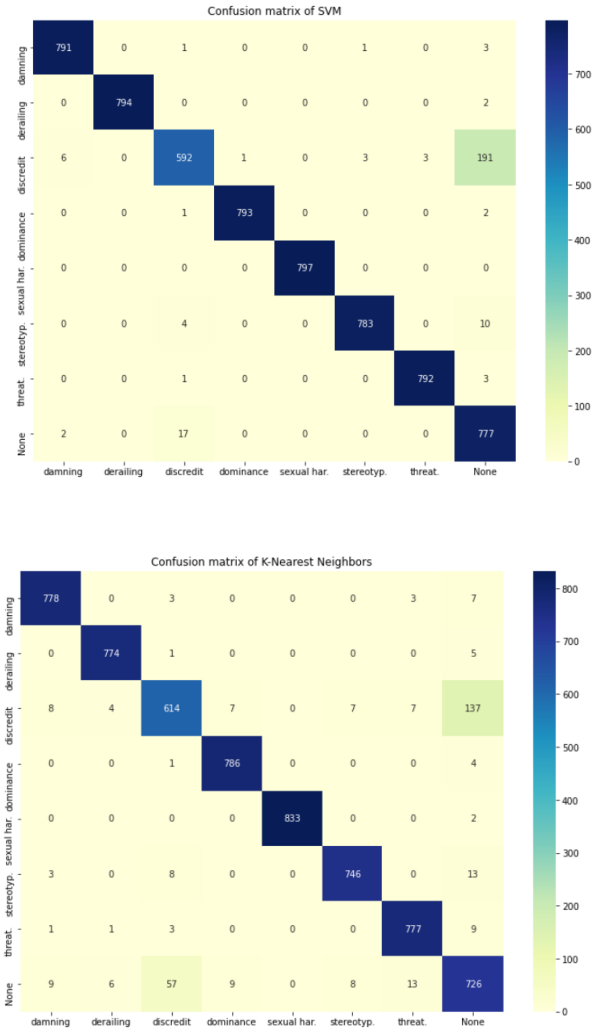
1) Bag of Words:

	Accuracy	Precision	Recall	F1
SVM	96.1%	0.97	0.96	0.96
kNN	94.7%	0.95	0.95	0.95
Decision Tree	91.5%	0.93	0.91	0.91

2) TF-IDF:

	Accuracy	Precision	Recall	F1
SVM	96.7%	0.97	0.97	0.97
kNN	93.8%	0.94	0.94	0.94
Decision Tree	92.1%	0.93	0.92	0.92

B. *Visualization of The Results*



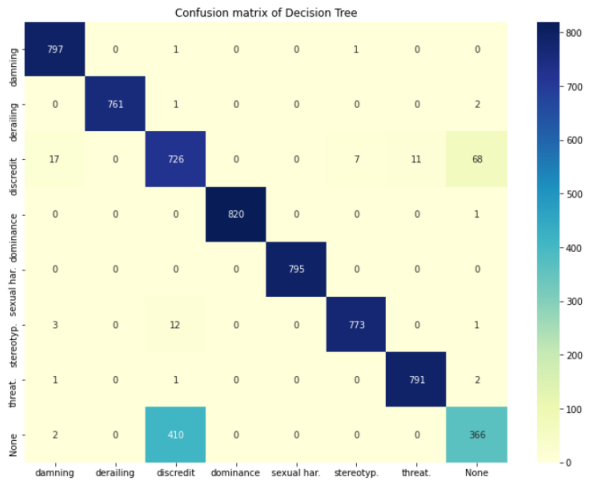
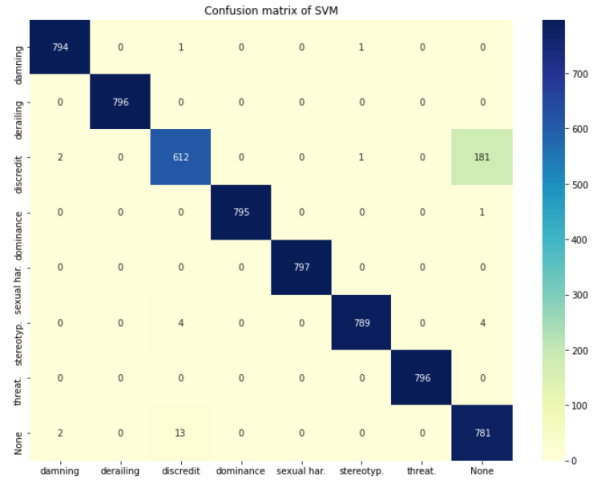
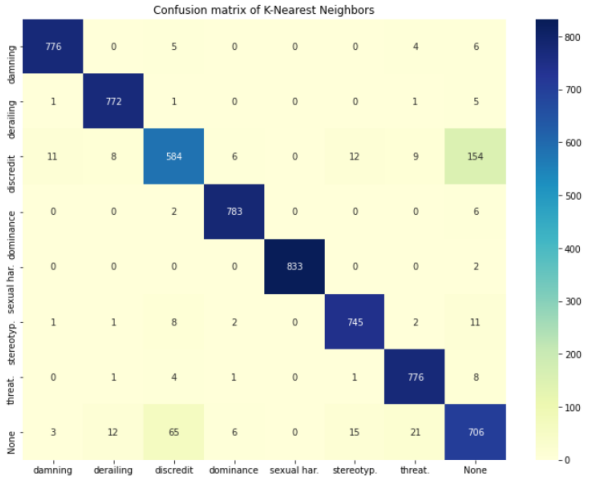


Fig. 6: Confusion Matrices of models using BOW





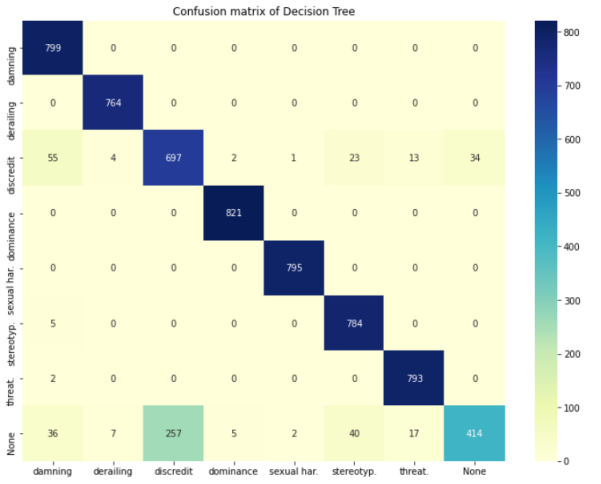


Fig. 7: Confusion Matrices of models using TF-IDF

C. Error Analysis

Given the previous results, we perform error analysis on the highest accuracy model which is SVM using TF-IDF transformation and we investigate the causes of the misclassified results.

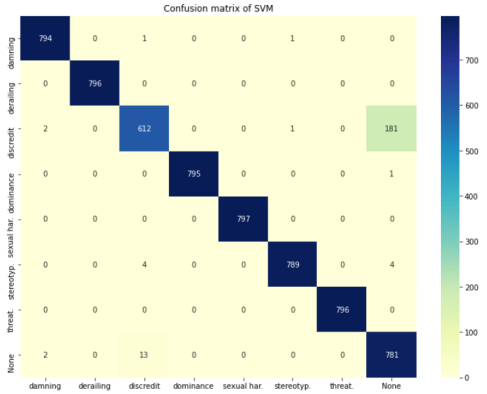


Fig. 8: Confusion matrix for SVM

1) **Discredit label** We notice that most miclassified cases are found when the true label is Discredit while the model predicts it as None. That happens in 181 instances. There are also 13 instances where the true label is None but the model predict it as discredit. This shows that the model is sometimes not able to detect the tweets where its intention is discredit. That is maybe because discredit sentences are not always direct and they hold hidden meanings.
2) **Mixed labels** Some tweets can belong to more than one misogyny category, but they are annotated according to the most dominant one. Example: a tweet where somebody insults a woman by saying (in translation to English): "You are a journalist! stay at home to cook food for others." This tweet is annotated as Stereotyping & Objectification where the man offends the woman. However, it can also be classified as discredit as he is discrediting her performance a s a journalist.
3) **Sarcasm** Model is not able to detect the sarcastic sense under which the tweets are written. Example: a tweet saying (in translation to English): "How you became a journalist?" This is annotated as Stereotyping & Objectification but because it looks like a normal question, the model is not able to correctly classify it.

IV. Clustering

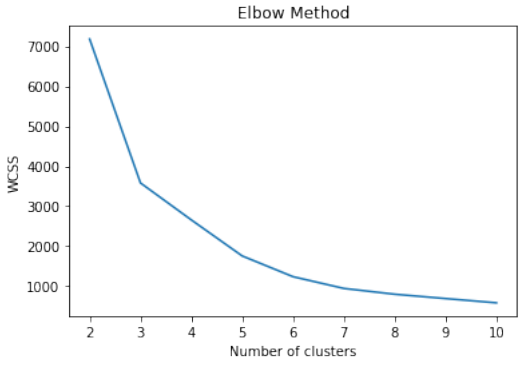
In the clustering step, we remove the label from the dataset and we let the model try to decide the cluster and then we visualize the model's cluster and its relation with the main labels of the dataset, using K-means model using two different feature engineering techniques: Bag of Words and TF-IDF.
We used PCA algorithm to reduce the dimensionality of the dataset due to the high dimensionality of the data so the training phase took a very long time .

A. Evaluation

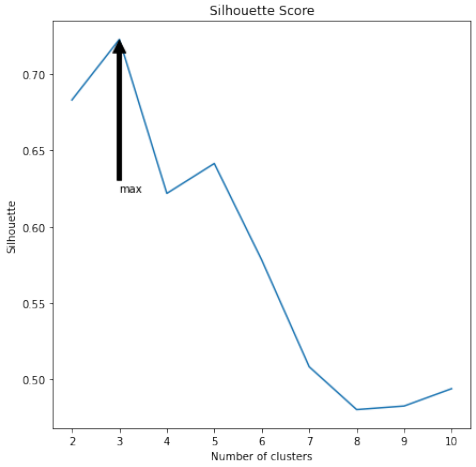
in the stage, we used the elbow method to know the number of clusters decided by the model .

1) Bag of Words:

- Elbow Methods:



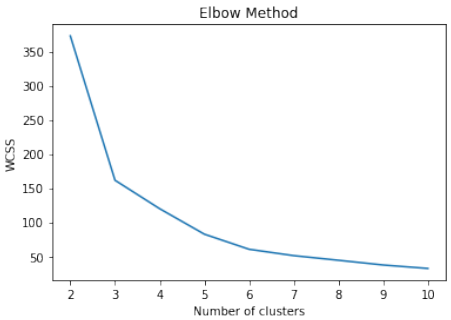
- Silhouette graph:
  From the graph we couldn't decide which k is the best, so we used the Silhouette as the evaluation method. So

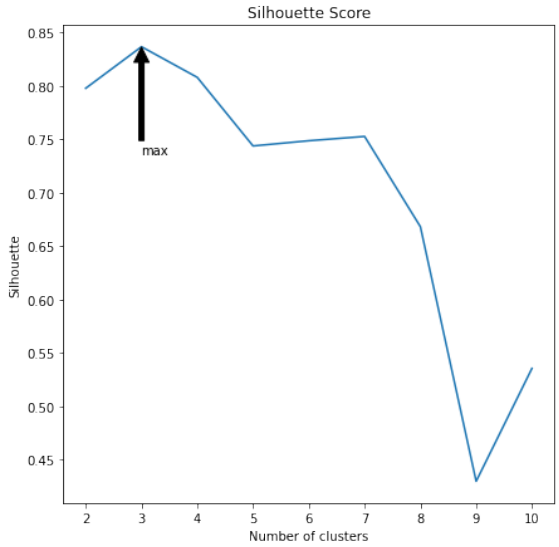


  in the graph we see that the best Silhouette at k=3 which means that the model divide all the dataset in only 3 classes.

2) TF-IDF:

- Elbow method:



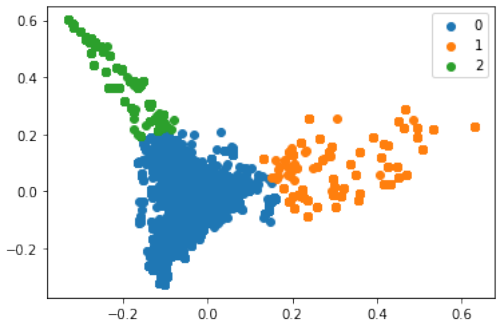


So in the graph we see that the best Silhouette at k=3 which means that the model divide all the dataset in only 3 classes.

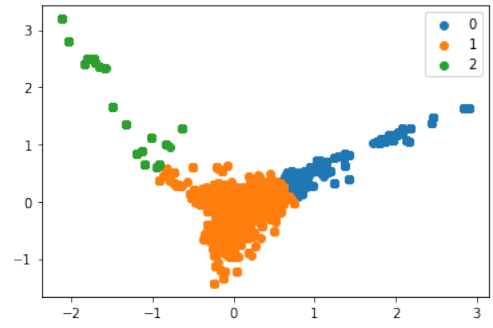
B. Visualization of The Results

In this section we will see how the model divide the dataset using the 2 different features engineering .

- Bag of Words:



- TF-IDF:



From the graphs we can see the plot of the pca dimension reduction algorithm using the two feature engineering and we can see how the model divide the 3 clusters on the dataset.

C. Error Analysis

In the final stage we will analyse the prediction of the model:

- the first cluster(cluster 0):
  many categories were label in cluster 0:
  ('discredit' 'none' 'damning' 'stereotyping and objectification' 'threat of violence' 'derailing' 'dominance' 'sexual harassment')
- the second cluster(cluster 1):
  many categories were label in cluster 0:
  ('discredit' 'sexual harassment')
- the third cluster(cluster 2):
  many categories were label in cluster 0:
  ('dominance' 'none' 'discredit')

Conclusion
We can see that the model couldn't detect each label alone , due to the similarity between the Arabic sentences, so in the first cluster nearly all the categories appeared in this cluster. for the second and third cluster, we can see that only 2 and 3 labels appeared in the clusters .

V. Transformers

The experiment was based on two transformers: mBERT [2] and XLM [3].

1) **Mbert** A pre-taining Model from BERT who supports 104 languages.The pre-training used The masked language modeling and the relation between sentences. His architecture contains 12-layer, 768-hidden, 12-heads, 110M parameters.
2) **XLM** The XLM model was proposed in Cross-lingual Language Model. It's a transformer pretrained using one of the two following objectives a causal language modeling (CLM) objective (next token prediction)or a masked language modeling (MLM) objective (BERT-like).

Fine tuning Experiment Result

	Accuracy	F1 Score	Matthews corr coeff
Mbert	96.57%	0.96	0.96
XLM	95.6%	0.95	0.95

Conclusion Mbert is the champion transformer with 96.57% accuracy percentage.

VI. Innovation

- Usually, the misogyny detection is made through classification. However, in our project we investigated the clustering approach as well, because in many cases the datasets available are not labeled. Thus, reaching a clustering model will be very helpful when more datasets are available.
- We do not want to only stop at the developer level detection. We would like our model to be deployed in an actual production environment, for further tuning, and

further improvement. Currently, the offensive tweets are handled manually. The platform users, see an offensive tweet, they report it, until an action from Twitter side is taken. Thus, we're reaching out to official Twitter Developer Platform, where Twitter team can listen to the feedback of developers around the world. We're offering them our idea that is, before a tweet is posted it gets through Twitter policy check which contains our model, and if it is a positive misogyny, the tweet is stopped and a warning message shows up to the user.

VII. Conclusion

As communication means keep developing, tools to regularize these means should also be developed, in order to make communication platforms a safe place for people.
Through data visualization we realized the intensity of the misogyny problem as we noticed the frequency and strength of offensive terms used.
We kept testing and exploring all possible approaches in order to reach an optimal performance. So, we used classification and reached SVM using TF-IDF transformation as a champion model. We also applied clustering and found out that the classes overlap due to the similarity in the language. In addition, we applied transformers and came up with Mbert as the best pre-training model.
For future work:

- The deployment of the model in an actual production environment and providing classification labels to user's inputs in this environments (Twitter itself for example).
- Further investigate the clustering approach using more datasets from the same dialect.
- Apply the pipeline to other Arabic Dialects.

Making the internet a misogyny free space is not a far away goal, with the contributions of social media platforms to provide actual users datasets, and the hard working data scientists together.

References

[1] H. Mulki and B. Ghanem, "Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language," in *Proceedings of the 6th Arabic Natural Language Processing Workshop (WANLP 2021)*, 2021.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.