King Saud University
College of Computer and Information Sciences
Information Technology Department

IT469
Human Language Technologies

hate speech Detection

Prepared by

Group#: 8	
Leader Email:438200309@student.ksu.edu.sa	
Ruba Alnashwan	438200309
Shaikha bin ateeq	438201519
Alanoud Alotaibi	437200539
Alanoud Alhamdan	438200089

Supervised by

<Afnan Al-Subaihin>
< Nora bin Madi>

ROLES AND RESPONSIBILITIES:

STUDENT NAME	ROLE	REVIEWED SECTION	CORRECTION MADE
Ruba Alnashwan	All team member	Text Cleaning + Preprocessing + Modeling	All team member
Shaikha bin ateeq	All team member	Introduction + Evaluation Results	All team member
Alanoud Alotaibi	All team member	Feature Engineering + data + format	All team member
Alanoud Alhamdan	All team member	Text Cleaning + Evaluation Results	All team member

Table of Contents

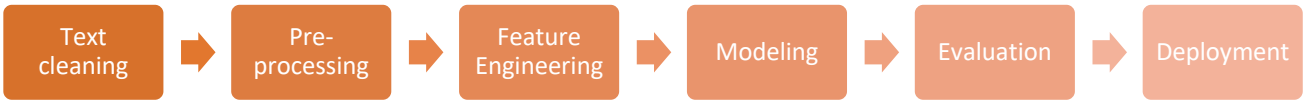
List of Tables

List of Figures

1. INTRODUCTION

People nowadays communicate through social media platforms from all over the world. However, verbal misbehaviors are now widely disseminated via social media. Twitter is one of the most well-known social media sites in the Arabic world.
To distinguish hate speech, its usually defined as "disparagement of an individual or a group on account of a group characteristic such as race, color, national origin, sex, disability, religion, or sexual orientation" such as "حيو" ,"*كل**" also Detecting hate speech in Arabic text, on the other hand, is still in its nascent, because of the diversity and complexities of the Arabic language's meaning and form.
We address this problem. To distinguish hate speech, Twitter's encouraged users to report any hate speech that violated its policies. To do effective action could be taken for the tweet. Despite Twitter's best attempts to combat hate speech, there are still tweets that endanger communities, and numerous people are still using it. It is significant because it is a social media site where many adolescents can follow and use certain words and acts and other things that could contribute to physical violence or deviation. We have an urgent need to discover and block such material on social media.
The project's goal is to create a high-end culture free of such words and behavior. We also recognize that technology can be a double-edged sword. We use its utility and technology to assist us, especially when it comes to Hate Speech Detection.

Our project will go through NLP pipeline which are:

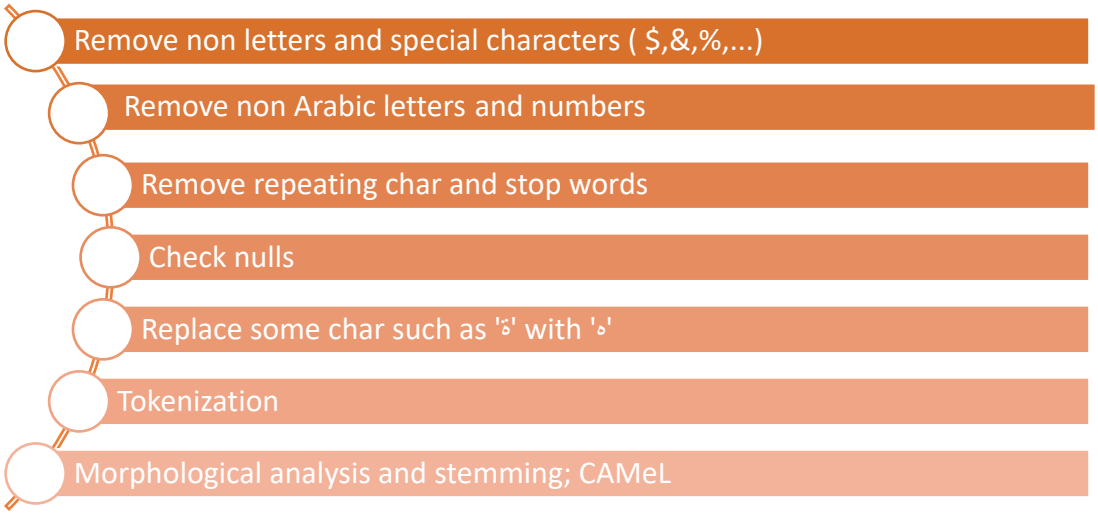


2. DATA

Our data set contains three columns which are tweet, Offensive, and Hate. The data set consists of training 6839 tweets and testing 1000 tweets for column Hate its provided two classes (NOT_HS, HS), and for column Offensive, it's provided two types (NOT_OFF, OFF). In our project, we want to detect hate tweets.

3. TEXT CLEANING AND PREPROCESSING

Preprocessing text data is an essential step because it makes the raw text ready for mining so that it becomes easier to extract information from the text and apply ML algorithms to it.
Before feeding the tweets as an input to the classification models, we applied several preprocessing steps:



According to Figure1 and Figure1, in this step, our goal was to remove noise that is less relevant to find the hate speech of the tweets, and we normalized the data to makes sure data reads the same way like (غريبه، غريبة) and to make data easy to handle by the classification models.

Figure 1: Dataset Before Cleaning.

	Unnamed: 0	Text	Offensive	Hate
0	0	الحمدلله يارب فوز مهم يا زمالك.. كل الدعم ليكم...	NOT_OFF	NOT_HS
1	1	فدوه يا بخت فدوه يا زمن واحد منكم يجيبه	NOT_OFF	NOT_HS
2	2	RT @USER: يا رب يا واحد يا أحد بحق يوم الاحد ا...	OFF	HS
3	3	RT @USER: #امي_يا وجع قلبي عليكي الحرية_هوا...	NOT_OFF	NOT_HS
4	4	🎼يا بكون بحياتك الأهم يا إما ما بدي أكون	NOT_OFF	NOT_HS
...	...	...	...	...
6834	6834	@USER يا حمار ، يا جاهل ، نسبة الباطل ما بتتحس...	OFF	NOT_HS

Figure 2: Dataset After Cleaning.

	Unnamed: 0	Text	Offensive	Hate	NHW
0	0	الحمدله يارب فوز مهم زمالك الدعم ليكم رجاله ...	0	0	0
1	1	فدوه بخت فدوه زمن واحد منكم يجيبه	0	0	0
2	2	رب واحد احد بحق يوم الاحد ان تهلك بني سعود ...	1	1	0
3	3	هوا الحريه وجع قلبي عليكي امي اله يحرق قلب...	0	0	0
4	4	بكون بحياتك الاهم بدي اكون	0	0	0

4. FEATURE ENGINEERING

Our project is about hate speech detection. It is well known that the models are strongly influenced by the quality and content of the training data. We collected our wordlist from two sources, which contain the total number of words that indicate hatred. Also, we have removed the repetitive and irrelevant word manually (Figure 3). We believe that when we focus on the tweets which contain hate words, the Accuracy will be better. On the other hand, the offensive phrases have misled us so much because there is no specific source for them, For example:

Figure 3: Hate Dataset.

```
[4] dfword=pd.read_excel("done.xlsx") #list of hate word
```

```
dfword
```

	term
0	لعنه
1	نجس
2	قرد
3	خنزير
4	كلب
...	...
395	مس
396	مشرك
397	مرض
398	مسيحي
399	مريض

After we got our wordlists, we started counting hate words in each tweets and we denote it with a number. We thought that the classification is dependent on the number of words of the hate and have seen in several papers their use of this method, but after calculating the number of hate words we found inconsistency with the classification. For example column, Hate represented the tweets as (HS, NOT_HS) and noticed there are 6 Hate Words in one of the tweets, classified as not hate and vice versa.

5. MODELING

We have used various modeling techniques such as TF-IDF, SVM, Logistic regression, and AraVec.

Logistic Regression: Logistic Regression supervised machine learning. It is a direct probability model.

Tf-idf: TF-IDF aims to clarify the importance of the word in the document. If the word appears several times in the text, it must be increasingly important. Simultaneously, if a word appears several times in a text, along with many other documents, it could be because that word is just a frequent word, not that it was necessary.

Support Vector Machine (SVM): This supervised machine learning in which we give data and classify them into two groups.

AraVec: AraVec is pre-trained in word embedding and is an effective word embedding model for the NLP community.

6. EVALUATION RESULTS

According to Table 1, we have used recall, precision, f1-score, and accuracy to measure the resulting models' performance.

Table 1: Evaluation Results.

Hate Speech Detection Using	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic Regression and Tf-idf	0.91	0.96	0.93
Logistic Regression and Tf-idf+ Other Features	0.96	0.96	0.94
(SVM) and Tf-idf	0.95	0.96	0.95
(SVM) and Tf-idf+ Other Features	0.96	0.96	0.96
Logistic Regression and Aravec	0.94	0.96	0.94
Logistic Regression and Aravec + Other Features	0.96	0.96	0.96
SVM and Aravec + Other Features	0.96	0.97	0.96

- Weighted avg precision ratio of predicted Positives is truly Positive which the highest result was (0.96) and that indicates the result was not bad.
- Weighted avg recall is the ratio of correctly predicted positive observations to all observations in actual class positive, which the highest result was (0.97), and that indicates the result was good.
- Weighted avg f1-score is a combined measure that assesses the P/R tradeoff which the highest result was (0.96), and that indicates the result was not bad.

By adding (Number of hate word) column to know how many hates word was on the text shows if the tweet was either hate or not hate. We used both Tf-idf and AraVec with supervised machine learning (SVM, Logistic regression), the SVM was better in both cases. But the better models for SVM was (SVM and Aravec + Other Features). We have used recall, precision, f1-score, and accuracy to measure the resulting models' performance.

7. REFERENCES

En.wikipedia.org. 2021. *Hate speech - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Hate_speech> [Accessed 24 March 2021].

Alshalan,Al-Khalifa, R., 2021. *Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach*. [online] Aclweb.org. Available at: <https://www.aclweb.org/anthology/2020.wanlp-1.2.pdf> [Accessed 24 March 2021].

Abuzayed, Elsayed, A., 2021. *Quick and Simple Approach for Detecting Hate Speech in Arabic Tweets*. [online] Aclweb.org. Available at: <https://www.aclweb.org/anthology/2020.osact-1.18.pdf> [Accessed 24 March 2021].

Wordlist1: GitHub. 2021. *raghadsh/Arabic-Hate-speech*. [online] Available at: <https://github.com/raghadsh/Arabic-Hate-speech/blob/master/Hate_pmi.tsv> [Accessed 24 March 2021].

Wordlist2: GitHub. 2021. *raghadsh/Arabic-Hate-speech*. [online] Available at: <https://github.com/raghadsh/Arabic-Hate-speech/blob/master/Hate_chi.tsv> [Accessed 24 March 2021].