

Arabic Dialect

By:

Batoul Alosaimi, Shroaq Almutiri and Norah Alqahtani



Introduction

The importance

Of this project comes from the importance of the Arabic language



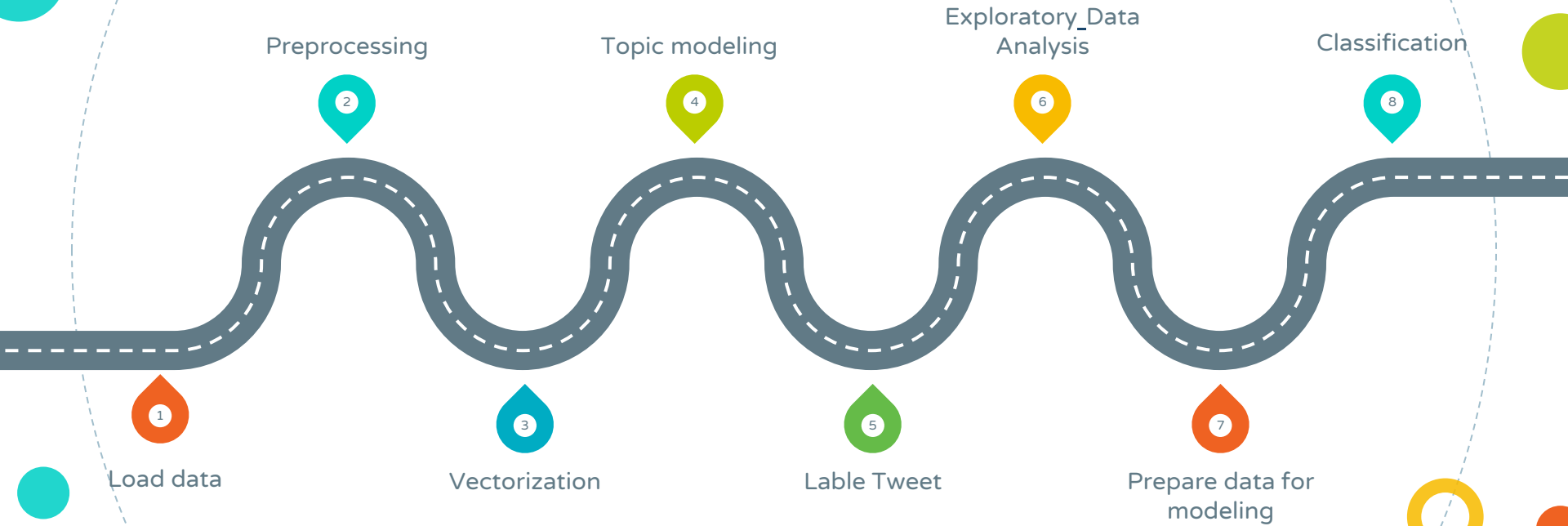
The objective is

to help institutions recognize and identify their customers in order to improve their services.

Dialect Maps



Methodology



Data Set



tweet	
11457	@Otty1986 ... يتشروط وتلو نصايح كإبكم دول متقدمة
5279	@doukha_azzedine ... الف مبرووووووك علينا ربي يحفظ
13104	@DerpinaJO ... طالع ابيض فيها يا دانا و راح اكون ب
25609	\n... مباح كنت بسوق تقريبا ل 5 ساعات ومعني صحابي
25831	...: لا آتحمّل مسؤولية أى حمار بيعط ويوقع بلساني
...	...
6551	@ToCjd @yyaraq1991 @6KAUbrTHli6LYqh @QZKRPr6VU...
10526	@nadooda8888 ... لا وكمان فتحوا له الكعبة دخل صلى
7736	@Amina68435787 @roroL47i ...ايوى صافي ما بقا لين
4355	...شكروا هوى تور لجبتوا البقرة مم من بين كل العر
26790	@maha_24i ...ما يسوى عليج 🙏 فلعوج مو سعوديہ ونادر

411949 rows × 2 columns

tweet	
11457	@Otty1986 ... يتشروط وتلو نصايح كإبكم دول متقدمة
25609	\n... مباح كنت بسوق تقريبا ل 5 ساعات ومعني صحابي
25831	...: لا آتحمّل مسؤولية أى حمار بيعط ويوقع بلساني
5172	...اصحي وخلي يكتسيدا ضي\nالصبح من نونك قل بيانه
8271	@maya_e2 ...مرض\والله تمام كل الامور تحت السيطرة
...	...
30170	...🙏 البندات ياربي 🙏 بصااa

62368 rows × 2 columns

Pre-processing

- Text cleaning like remove tashkeel , URLs, username, hashtags and spam word

- Remove repeated letters :

ياااااارب -> يارب

- Correct word:

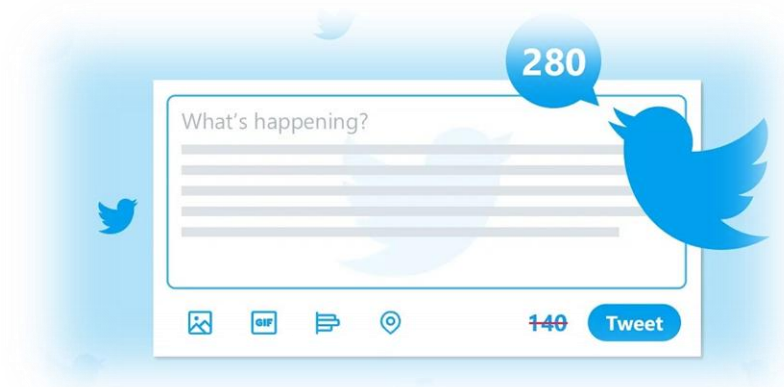
ف -> في

ع -> على

- Simplify :

أأأ -> أ

ة -> هـ





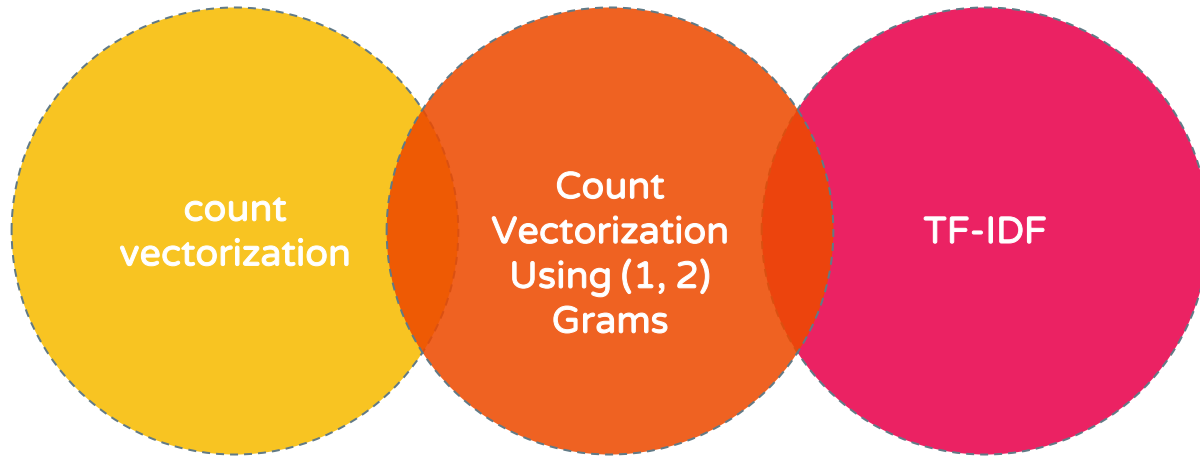
Remove stop word

- © We used a CSV file containing 750 Arabic stop words
- © Common words
- © add new stop word after topic model

word cloud



Vectorization



The background features a light gray dashed line forming a large circle. Various colored circles and arcs are scattered around: a large yellow-green circle at the top left, a small green circle with a white dot, a small blue circle, a large orange circle at the bottom left, a small pink circle, a large cyan arc at the top center, a medium blue circle containing the quote mark, a large yellow circle at the bottom right, a large orange arc, a small green circle with a white dot, and a small cyan circle.

“

Topic Modeling

- LSA (components=4)
- NMF (components=4)

LSA and TF-IDF Vectorizer is best models

Nilotic

اعتقلوه
حاجز
ماحدا
اخذوه
بيشتغل
تبع
بيعرف
مصطفى
عمرو
الرزق

Gulf

شنو
ليك
الليلة
فهموني
محتاره
الكياة
بيشوف
قولنا
دافعين
قصته

Levantine

حدا
بدي
منيح
لانو
متل
بدنا
نحنا
ضل
ريت
بعدك

Moroccan

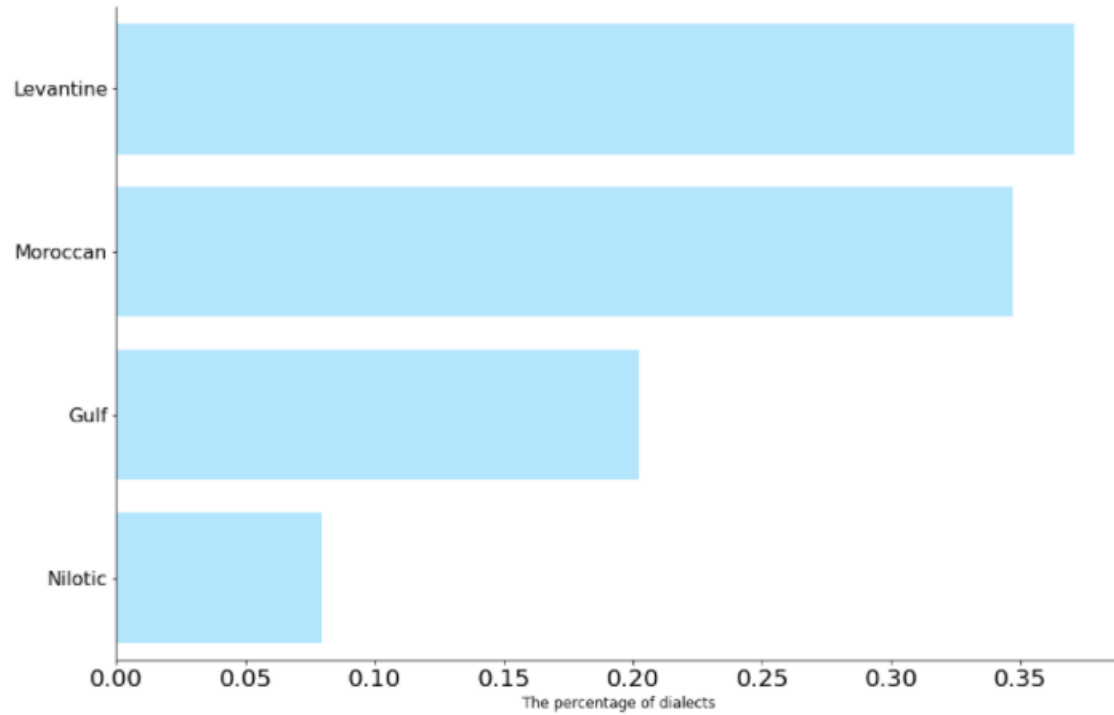
الحكي
ديال
بحال
بزااف
هون
باش
ليا
كلشي
بسمع
عايز

Label Tweet

	tweet	topics	Dialect	Nilotic	Gulf	Moroccan	Levantine
0	... بتشرطو وتندو نصايح كانتكم متقدمه وناجحين اوما	2	Moroccan	0.000083	0.000412	0.000504	0.000489
1	... مبارح بسوق ومعى صحابى فجاه شب بتطلع فىا ويقل	2	Moroccan	0.000970	0.001470	0.002382	0.002091
2	اتحمل مسؤوليه بيغلط ويوقع بلساني	3	Levantine	0.000058	0.000059	0.000188	0.000916
3	...قفل بيبانه اصحي وخلى يكتسبنا ضيه الكسلانه ويسع	2	Moroccan	0.000090	0.000139	0.000613	0.000125
4	السيطره مرضنا عتغير مرضه بسيطه كيفك اخبارك	3	Levantine	-0.000094	0.000548	0.001308	0.002030
...
63199	ياربى بصاير	1	Gulf	0.000209	0.002569	0.002393	-0.000157
63200	...بذك بروكسي بذك جيلبيريك لانو البرامج تبع البرو	0	Nilotic	0.016789	0.001105	0.010964	0.006855
63201	واشواقي غلا يف واهتويت اتعلم تهنى ريت	3	Levantine	0.000153	0.000335	0.001602	0.002778
63202	...نحاول نقرب بيحاولوا باى يبعدوا المسافات نهايتها	3	Levantine	0.000236	0.000308	0.000430	0.000724
63203	...ايوى صافي بقا فائز التناول والمواقع يعطيو رايه	2	Moroccan	0.000424	0.002221	0.002857	0.000508


60943 rows × 7 columns

Label Tweet



Classification model





	Accuracy	F1 score
Logistic Regression	0.7295	0.7295
Naive Bayes	0.6420	0.6420
SVM	0.7469	0.7469



Flask API

- ◎ Save the Model
- ◎ Create a function that takes the tweet and extracts the dialect type
- ◎ Build website using Flask API
<http://127.0.0.1:5000/>



Conclusion

- ◎ We analyzed 62,368 tweets
- ◎ We were able to create a model capable of analyzing Arabic texts and detecting their dialect with an accuracy of 75%.



Future work

- We aspire to improve the model results to the best results

- We scrape data from Twitter

Thanks!



Any questions?