أكاديمية سدايا
SDAIA Academy

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

KAPLAN
PROFESSIONAL

METIS®

Sentiment Classification for Arabic Texts

Trainee: Yathrib Alqahtani

Trainer: Prof. Essam Al Daoud

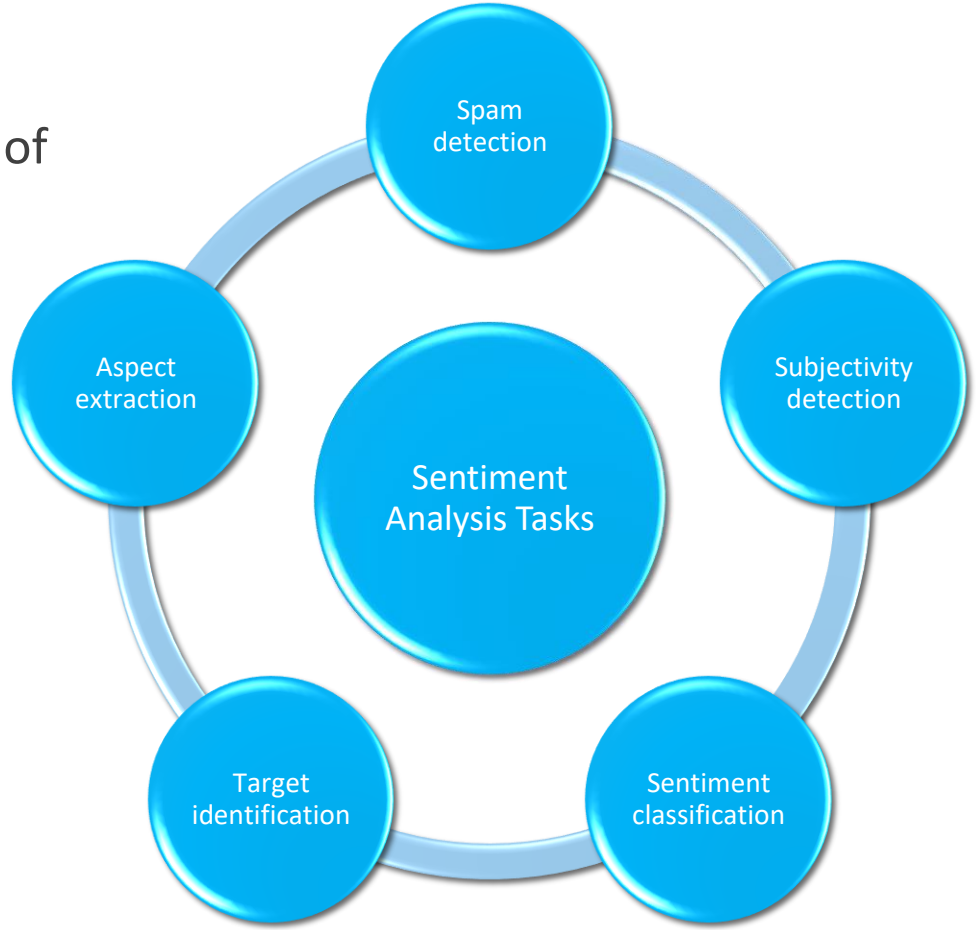
Outline

- Introduction
- Sentiment classification
- Datasets
- Cleaning and Preprocessing
- Modeling
- Results
- Conclusion

Introduction

- The rapid development and spread of Internet technologies has led to a dramatic increase in the number of Internet users viewing and creating content online.
  - A massive amount of opinionated documents.

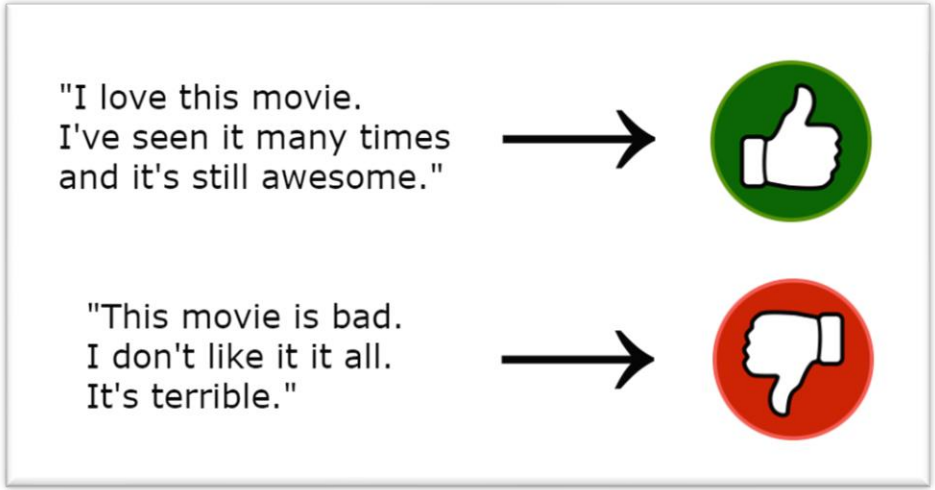
- The automatic extraction of opinions has became an active and broad research field in Natural Language Processing (NLP), referred to as sentiment analysis (SA)



B. Liu, *Sentiment analysis and opinion mining*, vol. 5, no. 1. Morgan & Claypool Publishers, 2012.

Sentiment Classification

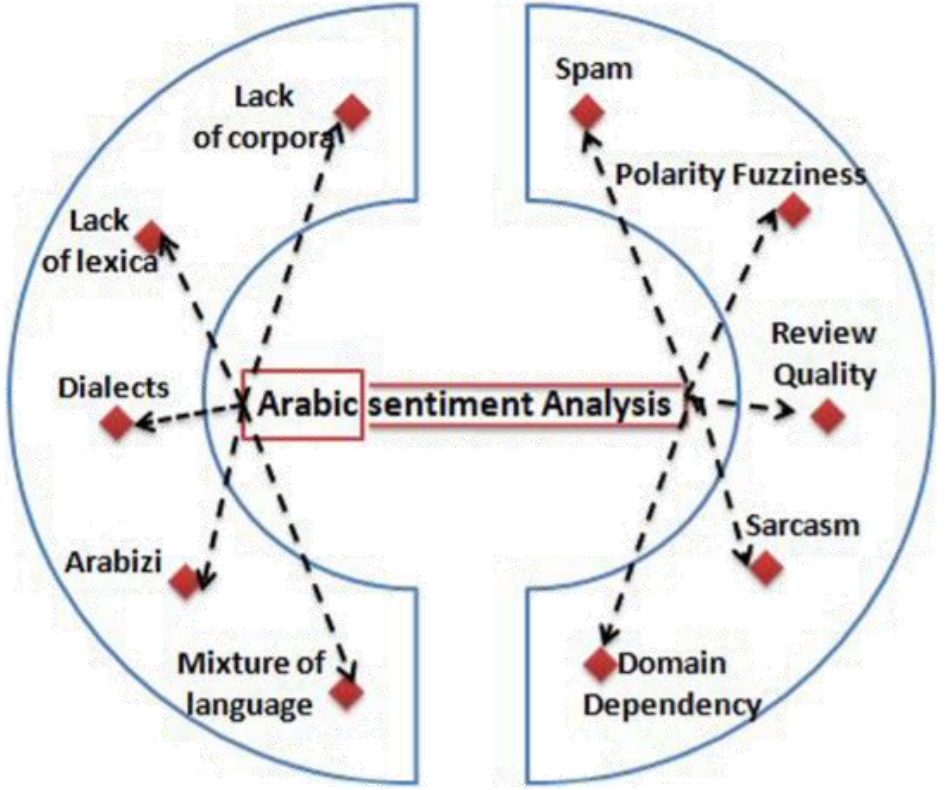
- Sentiment classification deals with the identification of sentiment as either positive or negative.
- Applications have reached almost all domains, from finance and healthcare to sports and politics.
- Recent supervised deep learning models have achieved considerable progress.



Sentiment Classification

- Arabic is the fourth most used Internet language after English, Chinese and Spanish.
- Around 185 million web users are Arabic speakers.

- Many Challenges of building models for the Arabic!



Oueslati, Oumaima, et al. "A review of sentiment analysis research in Arabic language." *Future Generation Computer Systems* 112 (2020): 408-430.

Sentiment Classification

- Examples of dialects variety in Arabic.

MSA word	Dialectical word	Arabizi	Country
حلو / جميل jameel	حلو	7elew	Lebanon
	حلو	7ilew	Saudi
	حلو	7low/ hlow	Tunisia
جدا jiddan	كتير	ktir	Lebanon
	وايد	wayed	Emirate
	أوي	2awi	Egypt
	برشا	barcha	Tunisia
دراجة Darraja	بسكلات	Besklet	Tunisia
	درا قة	Darraga	Egypt

Oueslati, Oumaima, et al. "A review of sentiment analysis research in Arabic language." *Future Generation Computer Systems* 112 (2020): 408-430.

Sentiment Classification

- Arabic vs English!

https://www.tarjama.com/arabic-dialects-different-types-of-arabic-language/

Dataset

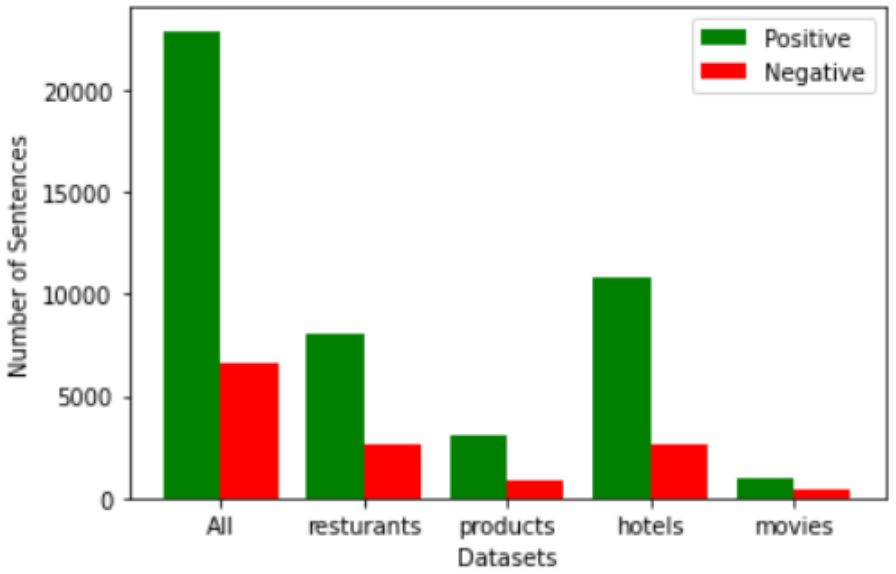
- **Large Multi-Domain Resources for Arabic Sentiment Analysis**
  - Modern Standard Arabic (MSA)
  - 29444 reviews
  - Restaurants, products, hotels, movies.

- **ArSenTD-Lev (Arabic Sentiment Twitter Dataset for LEVantine dialect)**
  - Dialectal Arabic (Levantine)
  - 3115 Tweets
  - Personal, sports, politics, religion.

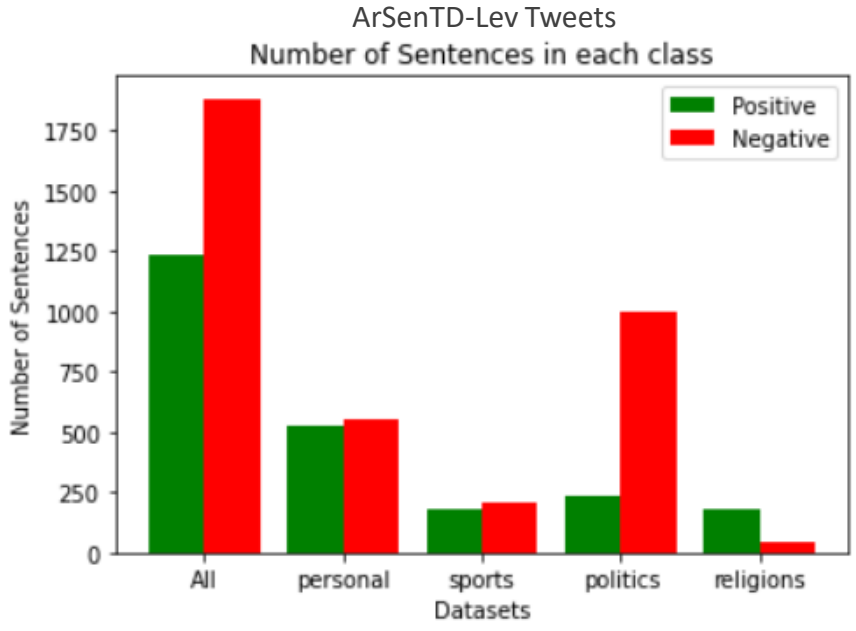
Dataset – Class distributions

- **Imbalance**
  - F1-socre is better than accuracy
  - Majority differs between reviews and tweets!

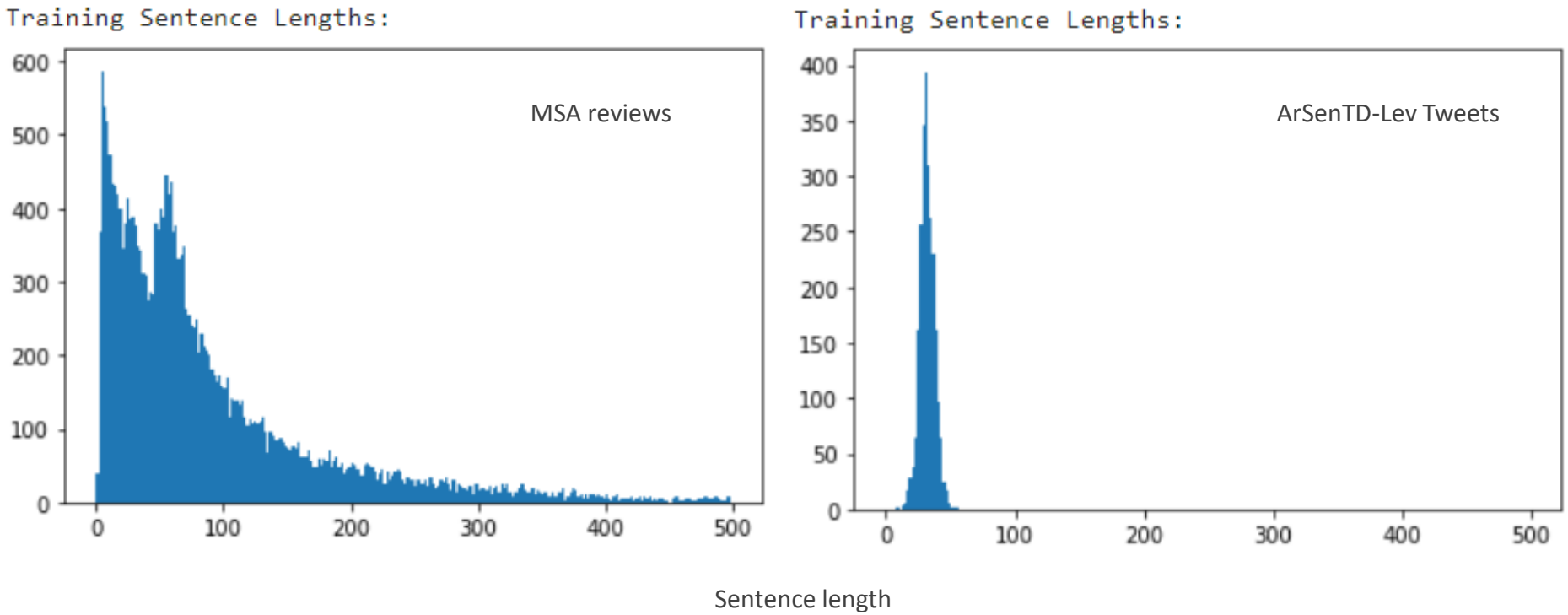
MSA reviews



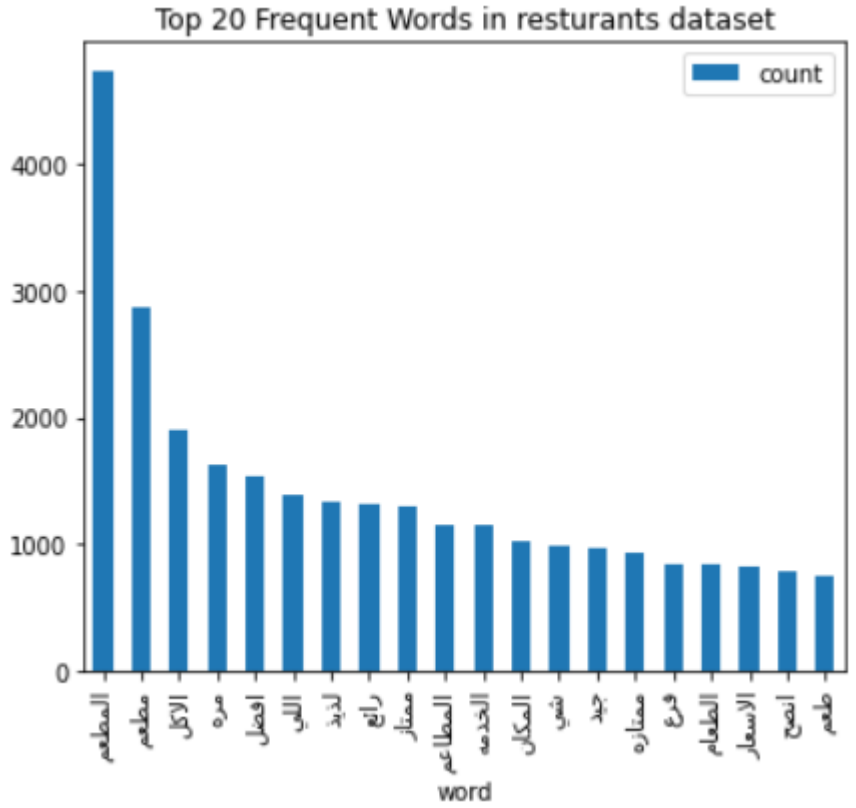
ArSenTD-Lev Tweets

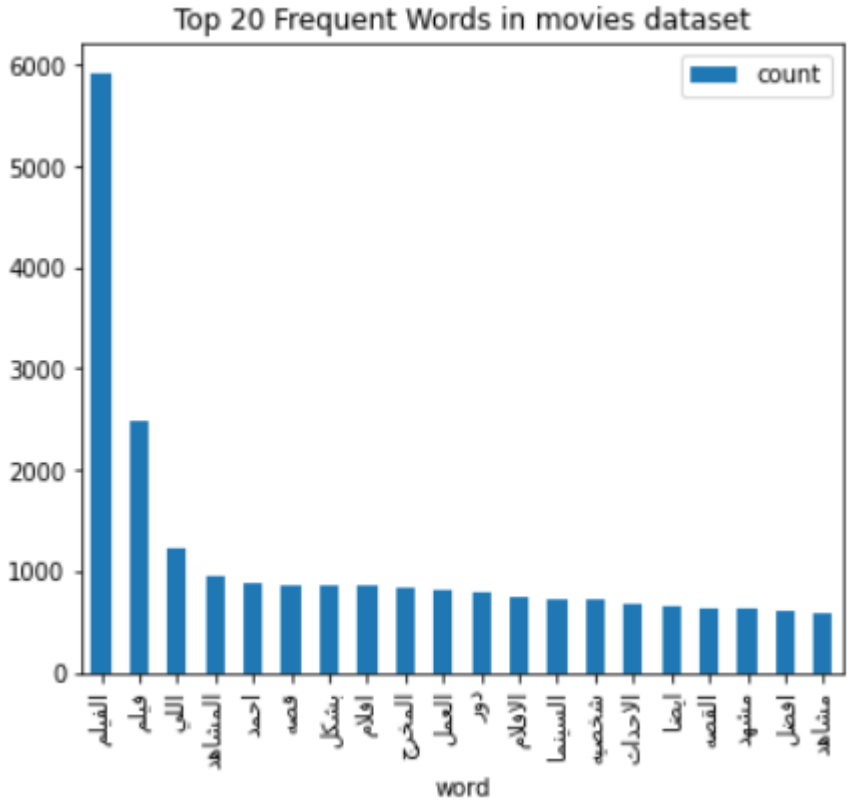


Dataset – Sentence lengths

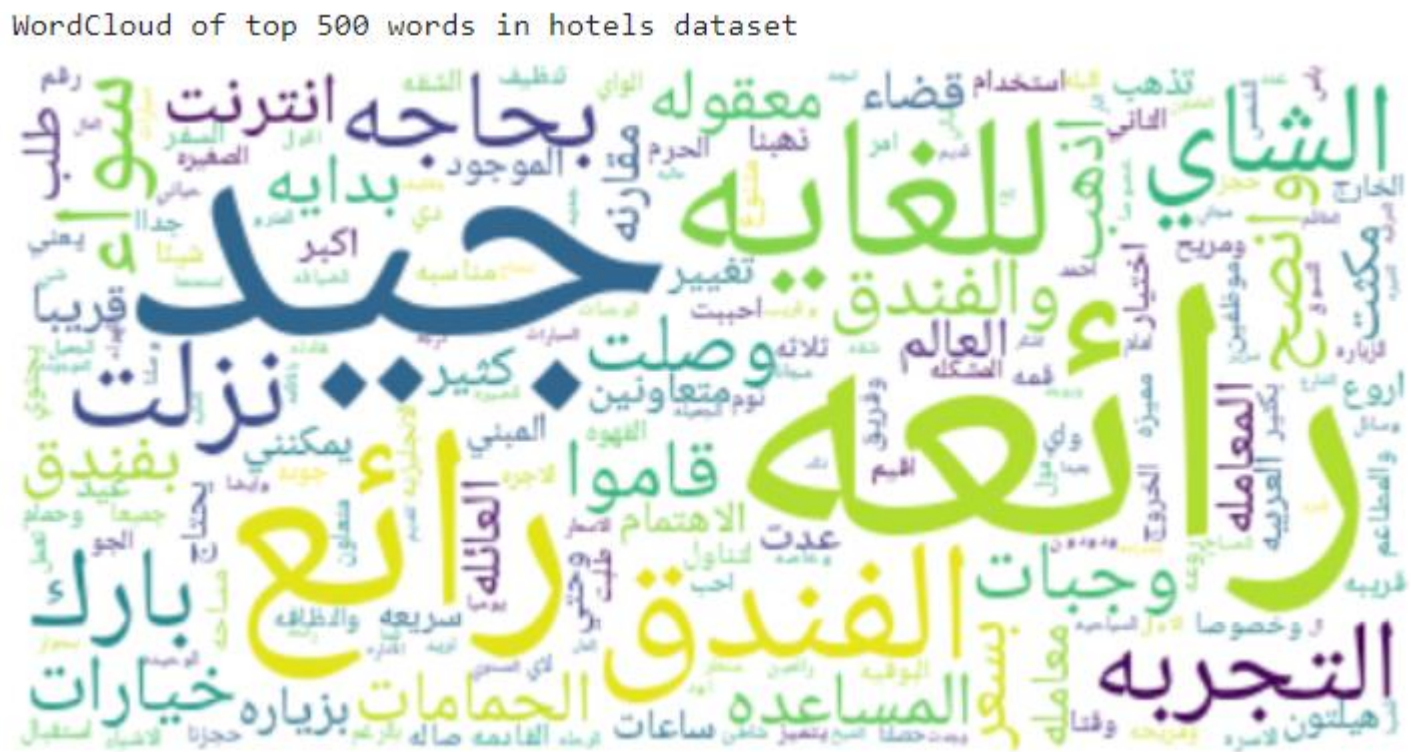


Dataset – Frequent features (MSA)

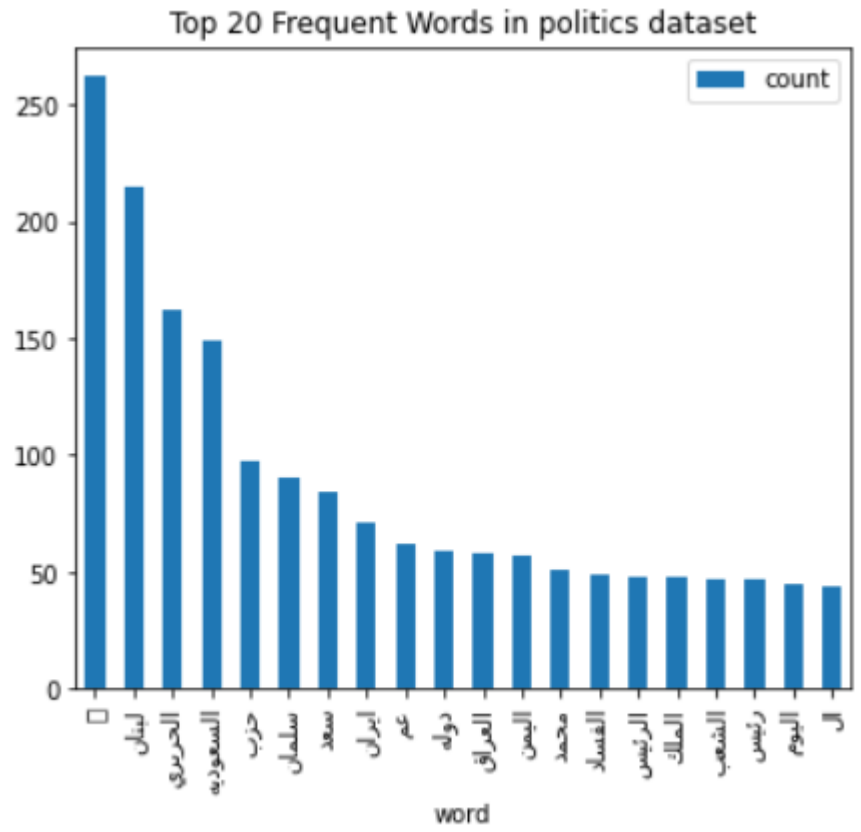


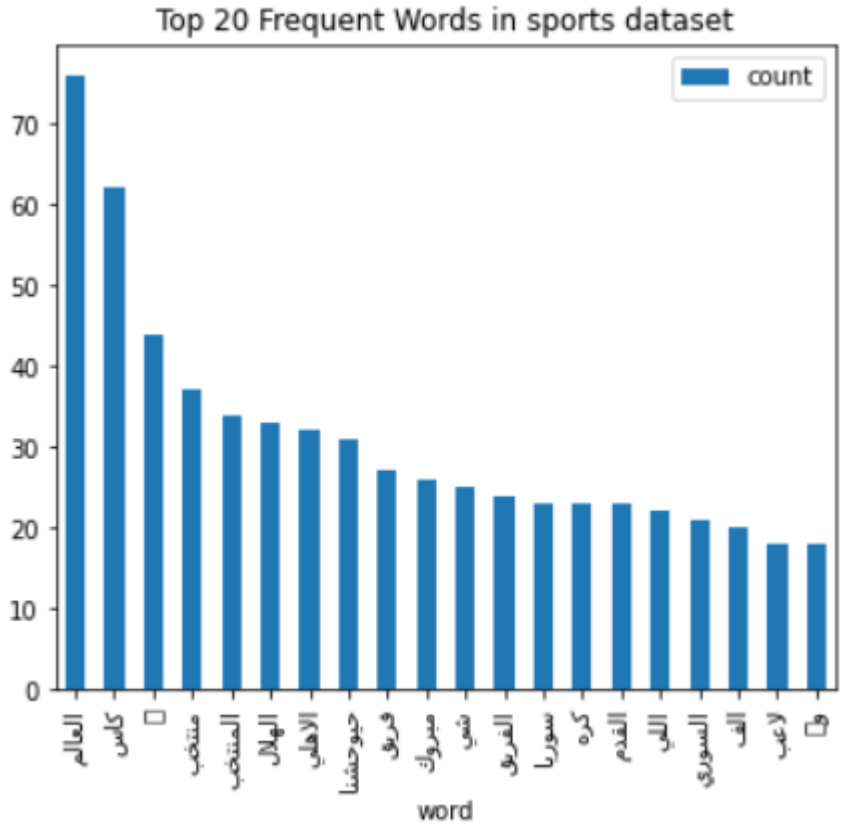


Dataset - Frequent features (MSA)

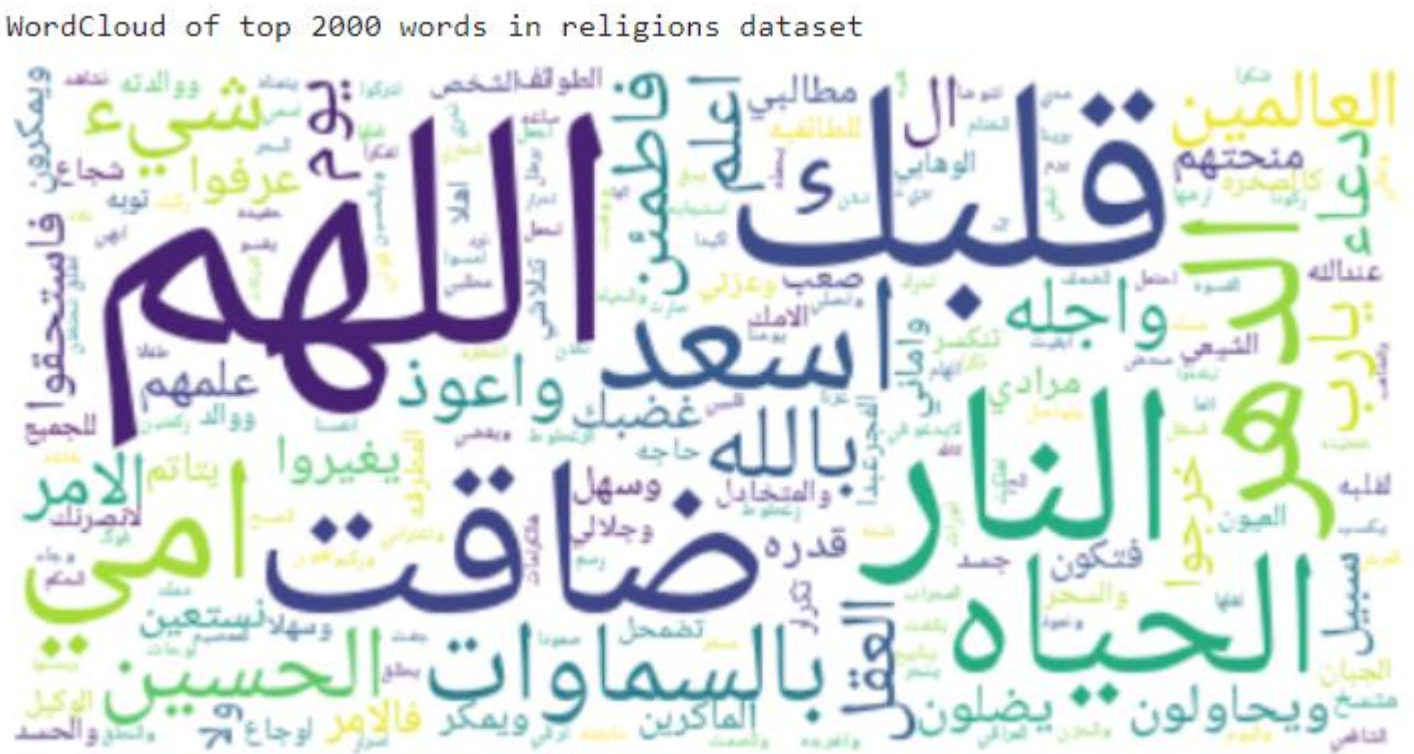


Dataset - Frequent features (ArSenTD-Lev)





Dataset - Frequent features (ArSenTD-Lev)



Cleaning and pre-processing

- Remove unwanted parts
  - Tashkel, URLs, punctuations, emojis, numbers and English letters, twitter hashtags and mentions.
- Remove Arabic stop words.
- Normalization.
- Stemming was tested but didn’t show improvements.

Cleaning and pre-processing

3@ ...منFInQe . عندما اعتقد كريستيانو انه ثاني افضل لاعب في العالم بعد كاكا و ميسي ثالثاً ثم حدث العكس 😂😂😂😂😂😂 https://t.co/n3EZ0pX0X7
@Yousef_MUFC اكثر ما يزعجنا بعد مستوانا خارج ارضنا . هو تمثيل كونتي بعد فوزه علينا
الف الف مبروك للمنتخب السوري كسب احترام الجميع لعب بروح والعزيمه شكرا على كل شي قدمتوه على مدار التصفيات النهايه برغم الظروف #استراليا_سوريا
مترجع للمنتخب . حتى اصدقائي بالثانوي اللي كنت فكرهم رمادين او حتى شبيحة طلعوا بيكرهوا المنتخب و شمتانين .. لك هالكم واحد وعامل جمعور منين جي ؟
@yara__CR7 والله الاتنين غالين على . لانو تقريبا اول نادين حبيتهم . حتي بس روح ع سعودي جبت تي شيرت الهلال 😂😂
{ اليابان* . _جرحت الهلال_ . بـ { نيتيمورا } . _حرمت الهلال_ من . { ادواردو*
@sh0uqa2 @v_i_b_2 #ي اتمنى . القوة القوة لا بارك الله بضعف هذا جمهور خط النار لا تورط نفسك معهم .٢. كاس_العالم_بيوحشنا… https://t.co/Qjuyf9vztz
😽 اسنسيو خرافي بكلشي اليوم . ما عنده قتالية الدون, بس كتير بيحاول يقدم
- لو كان (نصف) مزيج بين الظاهرة وزيدان .. لإحتفلنا بالهدف قبل ان يستلم الكرة ، ولكن الغباء يلعب دور ! https://t.co/jrTPJJjs4h
#نادي_ضمك مباراة مهمّة جداً جداً بالنسبة للنادي وإدارتة وجمهورهـ الكريم . الفوز ولاتشي غير ا… https://t.co/ougMBGzZiN تقام عصر اليوم على ملعب
منتخب فرنسا حالياً منتخب جبّار بكل خانة تحصلك نجم ، اتوقع يسوون شي بالمونديال القادم
منتخب لبنان يسجل هدفه الأول في الدقيقة الواحدة والعشرين على منتخب كوريا الشمالية ببطولة كأس آسيا على ملعب بيروت البلدي
هذا منتخب الاقليات الذي يعتبر ان الموت فقط هو ما يستحقه الاكثريه كيف كنت تؤيد هذا الفريق يا سني سوري شكرا استراليا https://t.co/FFV7eW7jhM
❤️حتى لو خسرنا. المهم إنكن رفعتوا راسنا قدام العالم كلّو و الكلّ عرف مين المنتخب السوري و مين نسور قاسيون. #نسور_قاسيون شكراً إلكن على هالفرحة
في حزة العصر جاني رساله رسالتٌ مؤلمه يا الله تعيني. شي يخص الهلال وباقي احواله . إصابة إد… https://t.co/xKokrpvihn
@oki_q8e شوف ابن اختي شوف 🤔 https://t.co/ZmVMl4sOqA . هدف لا نراه كل يوم #LEITOT
الخليوي_يستاهل فوق ماهو شايل الفريق هالموسم على ظهره تبرع للجمهور بـ١٠٠ تذكره على حسابه .. واتحدى اذا في لاعب اجنب… https://t.co/Ke41GlvB1O
@iAbdullah_24 😂 😂 😂 😂 والله اشي بقهر انو كل هل الانتقادات ع بنزيما وكل اشي . واخر اشي زيدان بحكي انو هو المهاجم الافضل ذ .
الشقيق التأهل لكأس العالم في #روسيا_2018 . عشت في الاوطانAM نبارك من القلب❤️ لمنتخب #المغرب & للعلى عنوان. ملء كل جنا… https://t.co/d05NSSo1DZ
💔 ديربي_مدريد #ريال_مدريد شيئا فشيئا المرنيقي يودع الدوري الاسباني فارق النقاط مع المتصدر 10 نقاط .. اجساد تتحرك في الميدان دون هوية وروح#
@OBalah1998 بحرف فيار. احضر هادا الفيديو من واحد من داخل الإتحاد الإسباني بيتكلم بشكل واضح عن المهر التحكيمي لصال… https://t.co/U3BUxVgKKX
الحمد لله... ماقصرتوا ياانجوم الزعيم.. مباراة قوية اضاع فيها الزعيم فرص كثيرة وسهلة ولكنها تدل على رغبة قوية في الفوز… https://t.co/Hggj8BVjr6
الف الف مبروك للمنتخب السوري كسب احترام الجميع لعب بروح والعزيمه شكرا على كل شي قدمتوه على مدار التصفيات النهايه برغم الظروف #استراليا_سوريا
كاس_العالم_حيوحشنا_٥. الجمهور لن يتوقف أبداً و سيكون ضد سياسة تركي الشيك #الأهلي خط أحمر و لن نرضى بذلك ابداً و لن نتوقف#
@IRAQFA. ثالثا أنتم في كل موسم تجعلون الدوري العراقي أطول دوري في العالم وهو يتكون من 20فريق. فكيف اذا كان الدوري 30فريق. تحياتي لكم يا فاشلين
نسور_قاسيون. يلا ياااا رب بدنا غوول تاني يا سومة هالفريق بعزيمتو وشجاعتو بستاهل يكون بكأس العالم حارب كل الظروف قلوبنا معكو وانشالله حتربحو#
الشقيق التأهل لكأس العالم في #روسيا_2018 . عشت في الاوطانAM نبارك من القلب❤️ لمنتخب #المغرب & للعلى عنوان. ملء كل جنا… https://t.co/d05NSSo1DZ
🙏❤️ اللهم أنصر نيمار و باولينهيو والشباب المحترمة على تشيلي، اللهم تعادل بيرو و كولومبيا، اللهم ثبت خطى ميسي و ماسكيرانو والشباب الوردة
@iAbdullah_24 😂 😂 😂 😂 والله اشي بقهر انو كل هل الانتقادات ع بنزيما وكل اشي . واخر اشي زيدان بحكي انو هو المهاجم الافضل ذ .
تعلمو يا #لبنان شو ناقصنا لنصير زي مصر .. الاعب اللبناني مفكر حالو ميسي كتير علينا اسيا يا حرام ... لك نحن وين و المنتخبات الباقية وين
في حزة العصر جاني رساله رسالتٌ مؤلمه يا الله تعيني. شي يخص الهلال وباقي احواله . إصابة إد… https://t.co/xKokrpvihn
#صح : 👑💚 وقبل دخول فترة الحرة. 👑💚 : صح. 👑💚 : الحين تركي يقول ان المبلغ اودع تاريخ ٢٨-١. 👑💚 : كاس_العالم_حيوحشنا_٥ . 💚… https://t.
@BarcaArabNet ماعتقد برجع يسيو ع او يسيو عين يحتاج اكثر من ذالك يحتاج مرونة في ساقيه و قدمين

الذاكره اعتقد كريستيانو افضل لاعب العالم كاكا ميسي ثالثا حدث العكس
يزعجنا مستوانا خارج ارضنا تمثيل كونتي فوزه
الف الف مبروك للمنتخب السوري كسب احترام الجميع لعب بروح والعزيمه شكرا شي قدمتوه مدار التصفيات النهايه برغم الظروف استراليا سوريا
مترجع للمنتخب اصدقائي بالثانوي اللي كنت فكرهم رمادين شبيحه طلعوا بيكرهوا المنتخب شمتانين هالكم وعامل جمعور منين جي
والله الاتنين غالين لانو تقريبا اول نادين حبيتهم روح سعودي جبت شيرت الهلال
اليابان جرحت الهلال نيتيمورا حرمت الهلال ادواردو
كاس العالم بيوحشنا القوه القوه بارك الله بضعف جمهور خط النار تورط اتمنى ي…
اسنسيو خرافي بكلشي اليوم قتاليه الدون كتير بيحاول يقدم
نصف مزيج الظاهره وزيدان لاحتفلنا بالهدف يستلم الكره الغباء يلعب دور
…يتقام عصر اليوم ملعب نادي ضمك مباراه مهمه بالنسبه للنادي وادارته وجمهوره الكريم الفوز ولاتشي ا
منتخب فرنسا حاليا منتخب جبار خانه تحصلك نجم اتوقع يسوون شي بالمونديال القادم
منتخب لبنان يسجل هدفه الاول الدقيقه الواحده والعشرين منتخب كوريا الشماليه ببطوله كاس اسيا ملعب بيروت البلدي
منتخب الاقليات يعتبر الموت يستحقه الاكثريه كنت تؤيد الفريق سني سوري شكرا استراليا
خسرنا المهم انكن رفعتوا راسنا قدام العالم كلو الكل عرف مين المنتخب السوري مين نسور قاسيون نسور قاسيون شكرا الكن هالفرحه
…حزه العصر جاني رساله رسالت مؤلمه الله تعيني شي يخص الهلال وباقي احواله اصابه اد
هدف نراه يوم شوف اختي شوف
…الخليوي يستاهل ماهو شايل الفريق هالموسم ظهره تبرع للجمهور ب١٠٠ تذكره حسابه واتحدي لاعب اجنب
والله اشي بقهر انو الانتقادات بنزيما اشي واخر اشي زيدان بحكي انو المهاجم الافضل
…نبارك القلب لمنتخب المغرب الشقيق التاهل لكاس العالم روسيا عشت الاوطان للعلى عنوان ملء جنا
ديربي مدريد مدريد شيئا فشيئا المرنيقي يودع الدوري الاسباني فارق النقاط المتصدر نقاط اجساد تتحرك الميدان هويه وروح
…يحرف فيار احضر هادا الفيديو داخل الاتحاد الاسباني بيتكلم بشكل واضح المهر التحكيمي لصال
…الحمد لله ماقصرتوا ياانجوم الزعيم مباراه قويه اضاع الزعيم فرص كثيره وسهله تدل رغبه قويه الفوز
الف الف مبروك للمنتخب السوري كسب احترام الجميع لعب بروح والعزيمه شكرا شي قدمتوه مدار التصفيات النهايه برغم الظروف استراليا سوريا
كاس العالم حيوحشنا الجمهور يتوقف ابدا سيكون ضد سياسه تركي الشيك الاهلي خط احمر نرضى ابداو نتوقف
ثالثا موسم تجعلون الدوري العراقي اطول دوري العالم يتكون فريق الدوري فريق تحياتي فاشلين
نسور قاسيون يلا يا بدنا غوول تاني سومه هالفريق بعزيمتو وشجاعتو بستاهل يكون بكاس العالم حارب الظروف قلوبنا معكو وانشالله حتربحو
…نبارك القلب لمنتخب المغرب الشقيق التاهل لكاس العالم روسيا عشت الاوطان للعلى عنوان ملء جنا
اللهم انصر نيمار باولينهيو والشباب المحترمه تشيلي، اللهم تعادل بيرو كولومبيا، اللهم ثبت خطى ميسي ماسكيرانو والشباب الورده
والله اشي بقهر انو الانتقادات بنزيما اشي واخر اشي زيدان بحكي انو المهاجم الافضل
تعلمو لبنان شو ناقصنا لنصير زي مصر الاعب اللبناني مفكر حالو ميسي كتير اسيا حرام وين المنتخبات الباقيه وين
…حزه العصر جاني رساله رسالت مؤلمه الله تعيني شي يخص الهلال وباقي احواله اصابه اد
كاس العالم حيوحشنا الحين تركي يقول المبلغ اودع تاريخ ٢٨ صح دخول فتره الحره صح

Modelling

Linear SVM	TF-IDF with unigrams and bigrams
AraBERT	AraBERT word embeddings
Bi-LSTM	Arabic word2vec
Bi-LSTM	Keres tokenizing and padding (indexing)

- Linear SVM
  - Default parameters
- AraBERT
  - Version v02
- Bi-LSTM
  - MSA reviews → a version pretrained on Wikipedia
  - ArSenTD-Lev tweets → a version pretrained on Tweets

- **Tools:** Google Collab for cloud processing / Pretrained language model (Hugging face) / Pretrained words embeddings (GitHub)

Results

Dataset	Domain	Size	SVM		AraBERT		Bi-LSTM-W2V		Bi-LSTM tokenizing/padding	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc
MSA	Restaurants	10705	0.92	0.87	0.94	0.91	0.90	0.84	0.90	0.85
	Products	3964	0.91	0.86	0.94	0.90	0.91	0.85	0.89	0.82
	Hotels	13422	0.97	0.95	0.99	0.98	0.92	0.87	0.95	0.93
	Movies	1353	0.89	0.83	0.90	0.85	0.84	0.71	0.83	0.72
	All	29444	0.94	0.90	0.96	0.94	0.59	0.49	0.93	0.88
ArSenTD-Lev	personal	1078	0.79	0.79						
	Sports	384	0.82	0.83						
	Politics	1237	0.67	0.90						
	Religions	227	0.92	0.87						
	All	3115	0.76	0.83	0.87	0.90	0.82	0.87	0.75	0.81

- AraBERT gave the best results in MSA and dialects and in all domains.

Results

Dataset	Domain	Size	SVM		AraBERT		Bi-LSTM-W2V		Bi-LSTM tokenizing/padding	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc
MSA	Restaurants	10705	0.92	0.87	0.94	0.91	0.90	0.84	0.90	0.85
	Products	3964	0.91	0.86	0.94	0.90	0.91	0.85	0.89	0.82
	Hotels	13422	0.97	0.95	0.99	0.98	0.92	0.87	0.95	0.93
	Movies	1353	0.89	0.83	0.90	0.85	0.84	0.71	0.83	0.72
	All	29444	0.94	0.90	0.96	0.94	0.59	0.49	0.93	0.88
ArSenTD-Lev	personal	1078	0.79	0.79						
	Sports	384	0.82	0.83						
	Politics	1237	0.67	0.90						
	Religions	227	0.92	0.87						
	All	3115	**0.76**	0.83	0.87	0.90	**0.82**	0.87	0.75	0.81

- Bi-LSTM-w2v didn't outperform simple SVM's in MSA reviews ,but it increased f1 score in ArSenTD-Lev tweets from 0.76 to 0.82

Results

Dataset	Domain	Size	SVM		AraBERT		Bi-LSTM-W2V		Bi-LSTM tokenizing/padding	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc
MSA	Restaurants	10705	0.92	0.87	0.94	0.91	0.90	0.84	0.90	0.85
	Products	3964	0.91	0.86	0.94	0.90	0.91	0.85	0.89	0.82
	Hotels	13422	0.97	0.95	0.99	0.98	0.92	0.87	0.95	0.93
	Movies	1353	0.89	0.83	0.90	0.85	0.84	0.71	0.83	0.72
	All	29444	0.94	0.90	0.96	0.94	0.59	0.49	0.93	0.88
ArSenTD-Lev	personal	1078	0.79	0.79						
	Sports	384	0.82	0.83						
	Politics	1237	0.67	0.90						
	Religions	227	0.92	0.87						
	All	3115	0.76	0.83	0.87	0.90	0.82	0.87	0.75	0.81

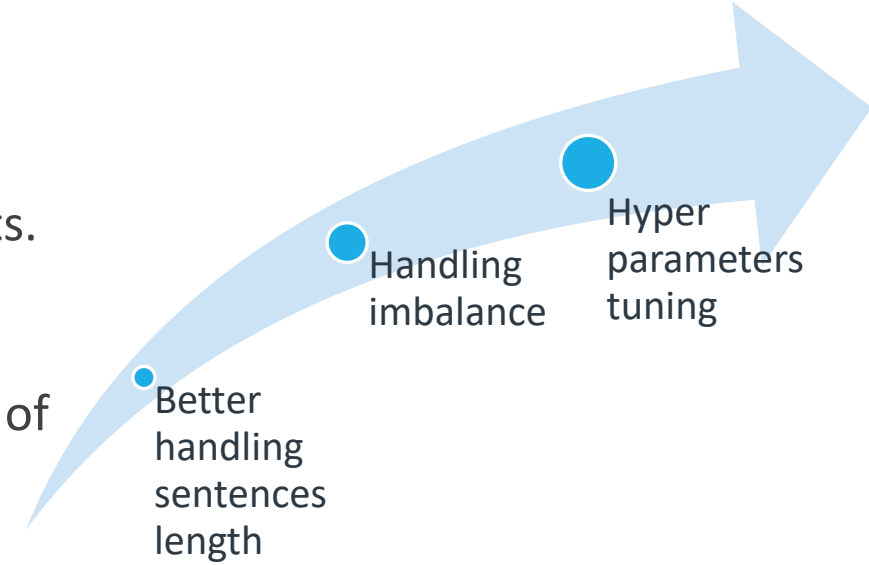
- Bi-LSTM-tokenizing/padding tends to beat or perform better than Bi-LSTM-w2v only for large datasets.

Conclusion

- This project focus on binary sentiment classification on Arabic texts.
  - MSA and Dialects
  - Reviews and tweets
  - Different domains

- There is substantial room for improvement for the reporting results.

- Extra efforts to be done to overcome challenges caused by variety of dialects in Arabic and limited availability of its resources.



Thanks for listening