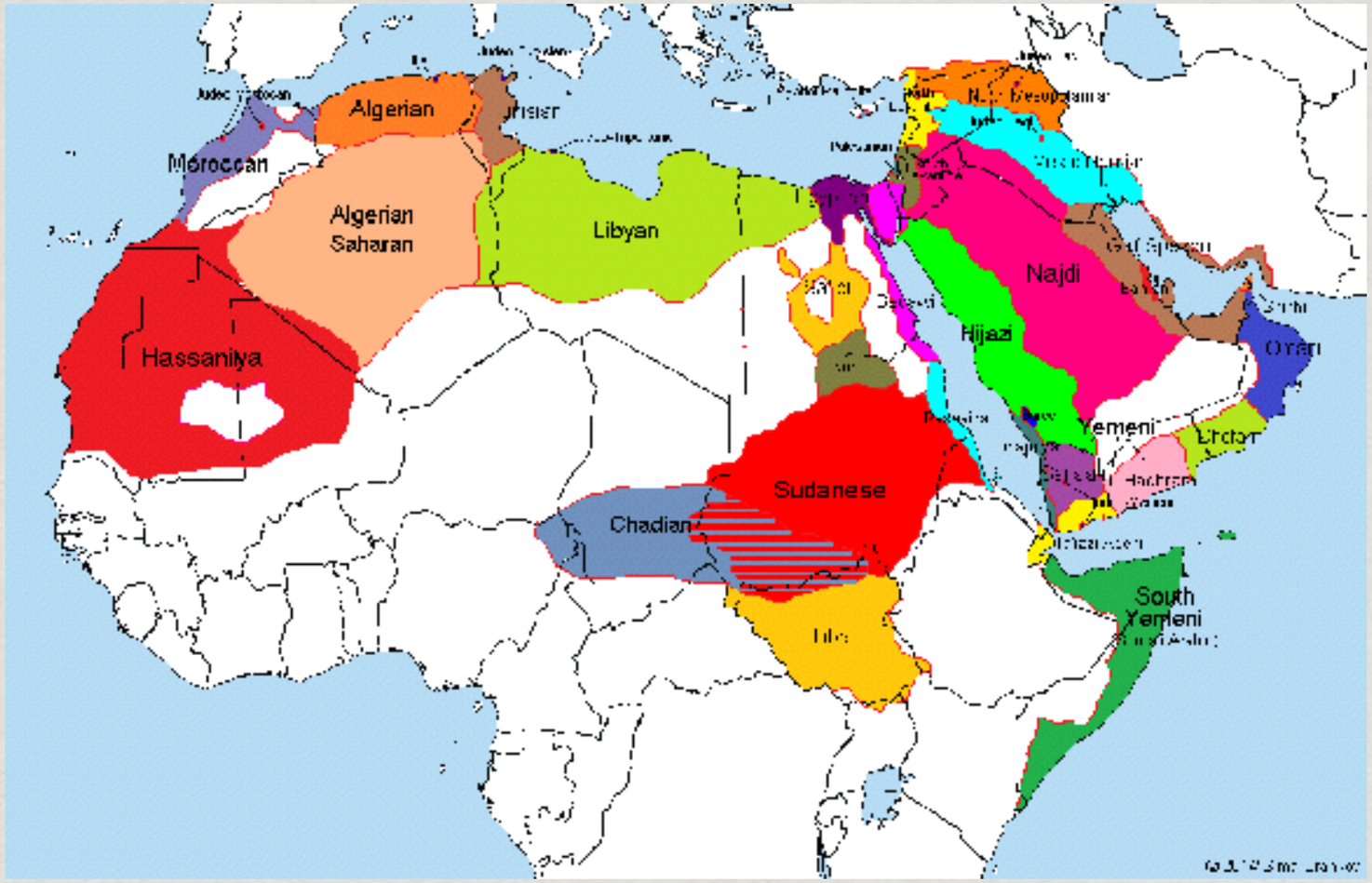
BUILDING DIALECT CLASSIFIERS USING TWITTER DATA

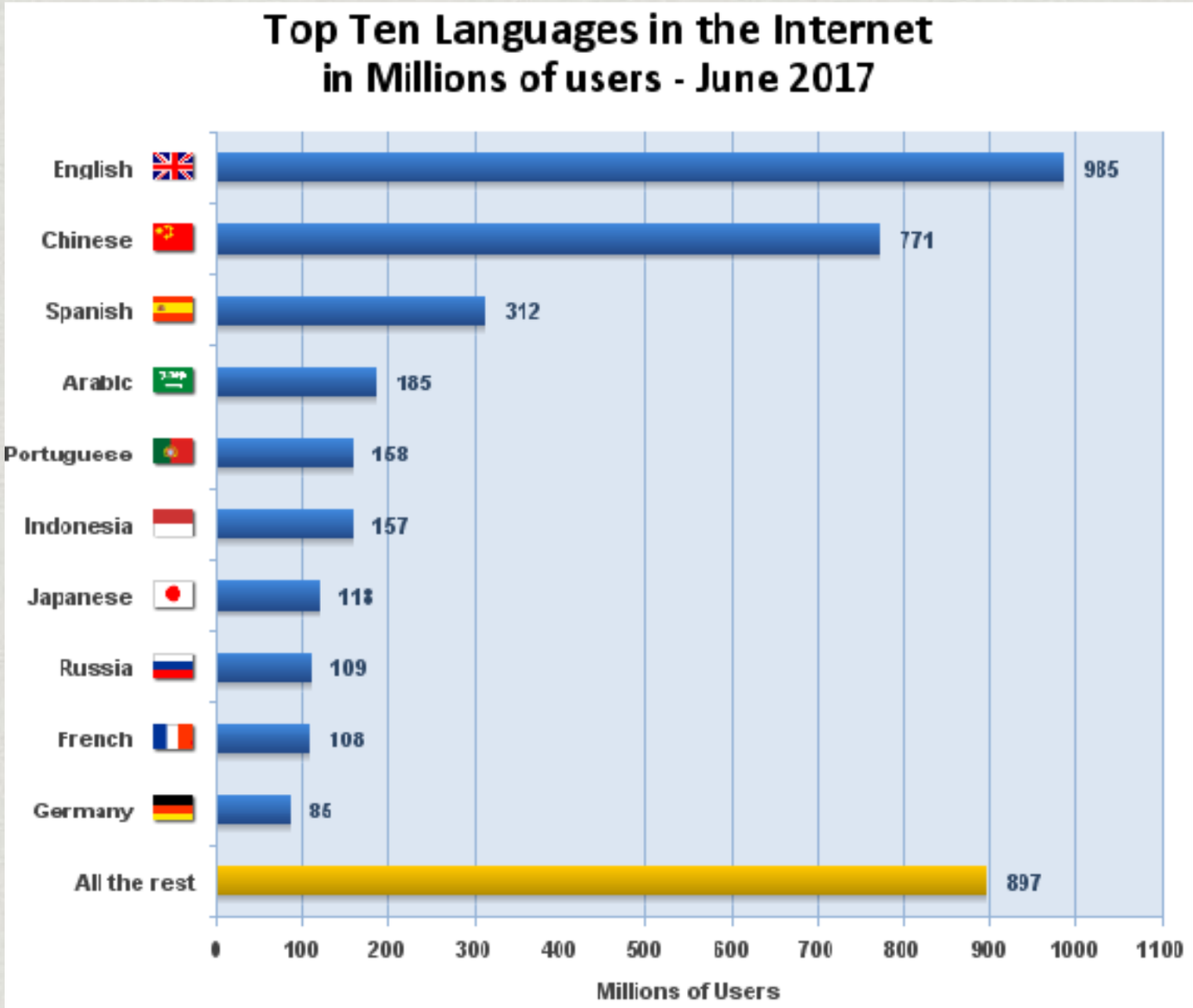
BY TAMIR ELSAHY

OVERVIEW

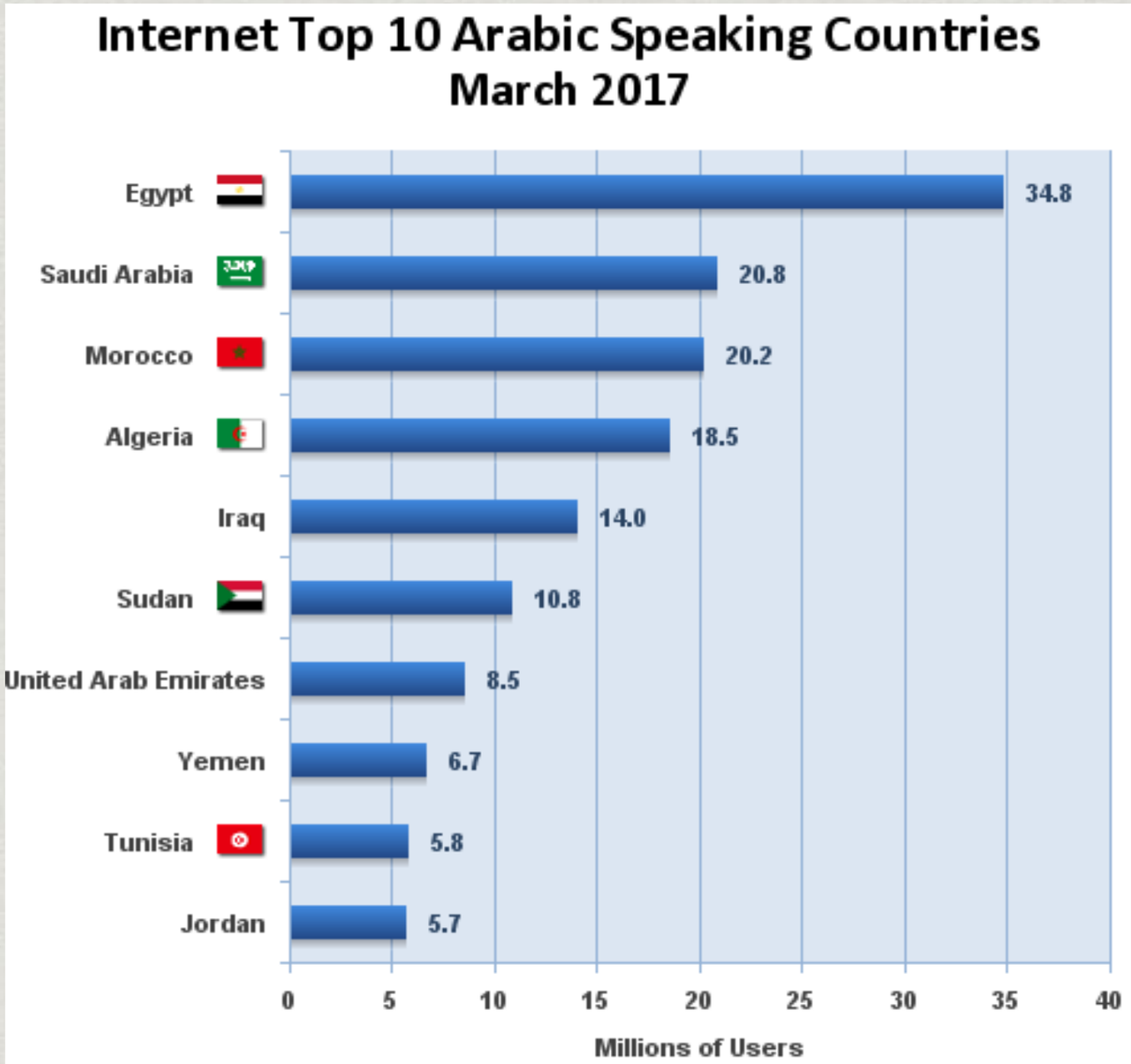
Arabic is the largest member of the Semitic language family and is spoken by nearly 500 million people worldwide. It is one of the six official UN languages. Despite its cultural, religious, and political significance, Arabic has received comparatively little attention in modern computational linguistics. Social media, like reader commentary on online newspapers, is a rich source of dialectal Arabic.



OVERVIEW


Top Ten Languages in the Internet
in Millions of users - June 2017
English
985
Chinese
771
Spanish
312
Arabic
185
Portuguese
158
Indonesia
157
Japanese
118
Russia
109
French
108
Germany
85
All the rest
897
0
100
200
300
400
500
600
700
800
900
1000
1100
Millions of Users
Source: InternetWorldStats - www.internetworldstats.com/stats7.htm
Estimated total Internet users are 3,835,567,619 in June 30, 2017
Copyright © 2017, Miniwatts Marketing Group

OVERVIEW



Source: Internet World Stats - www.internetworldstats.com
175,718,653 estimated Arabic Speaking Internet Users for March 2017

APPROACH

- Build a corpus of mostly informal Egyptian and Gulf Arabic that is rich in dialectal content was collected via the Twitter API.
- Divided into sub corpora that have been manually annotated according to dialect class.
- Topic models were used to observe interesting linguistic features through EDA and visualizations.
- Labeled tweets were then used to train and evaluate classifiers for dialect identification.
- Using this approach, the goal is to develop classifiers that significantly outperform baselines that use large amounts of MSA (Modern Standard Arabic) data.

QUERY SEARCH

- Use region specific dialect keywords to stream tweets:

    I. GULF terms for Hijazi, Najadi, Omani, Kuwaiti, Emirati, etc.

    II. EG terms for Alexandrian, Sa'idi, Badawi, etc.

- Avoid polysemous words when querying the streamer and inspect tweets for homographs in order to minimize one dialect leaking into the other's dedicated stream before the classes have been properly annotated.

EG	GULF
السكه	الطريج
ازيك	خربز
ضوافر	ازبن
مناخير	ابخص
كوباية	شخاط
أزاي	البشكاره
تهييس	مغسلة
بتاع	ياهوي
برضو	تميلح
كدا	دوشق
شبه الفلوط	خاشوقه

TIMELINE SEARCH

- Identify users with high lexical diversity scores from our previous stream and add them to a new timeline search. That way we keep getting more of what we like, which are unique words per tweet and more tweets.

- Check if user has over 3200 tweets (not a must if user has good quality tweets that are heavy in dialectal terms and not too many retweets).

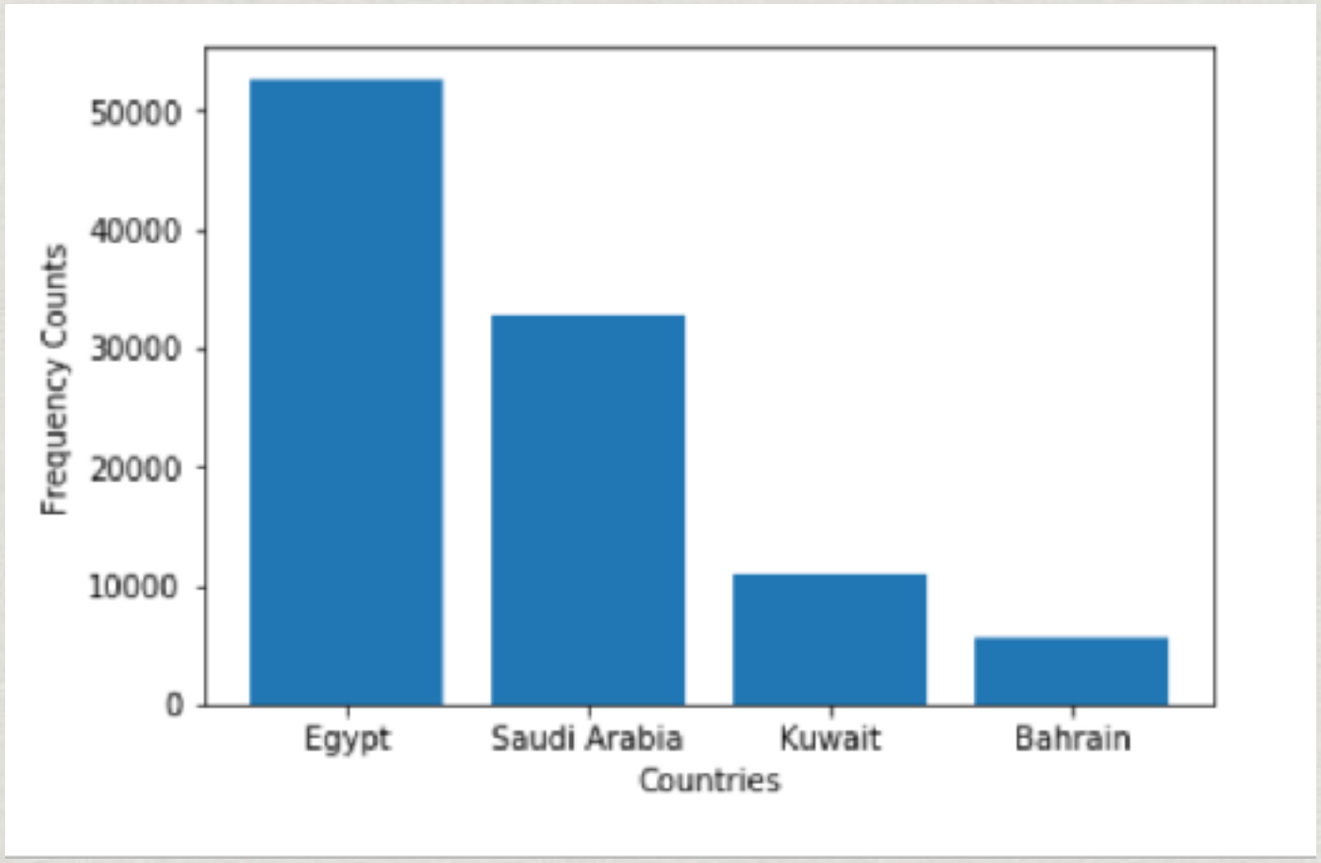
- Check for high volumes of polysemy and/or homographs.

- Prioritize timelines that focus on a specific category and find users whose tweets span as many categories as possible (fashion, politics, humor, parenting, romance, sensitive subjects, etc).

281	omjasem23	0.478261	لا لا حشا الطريج كله لك	23
296	hassanjasim	0.714286	توخو زحمة الطريج باجر	21
303	om_zeena_93	0.612903	بيسكرون الطريج عشان مؤتمر القمة	31
313	MHA25_	0.400000	لا راح اداوم بس انا على راسي ويشه يفتحون لي الطريج	50

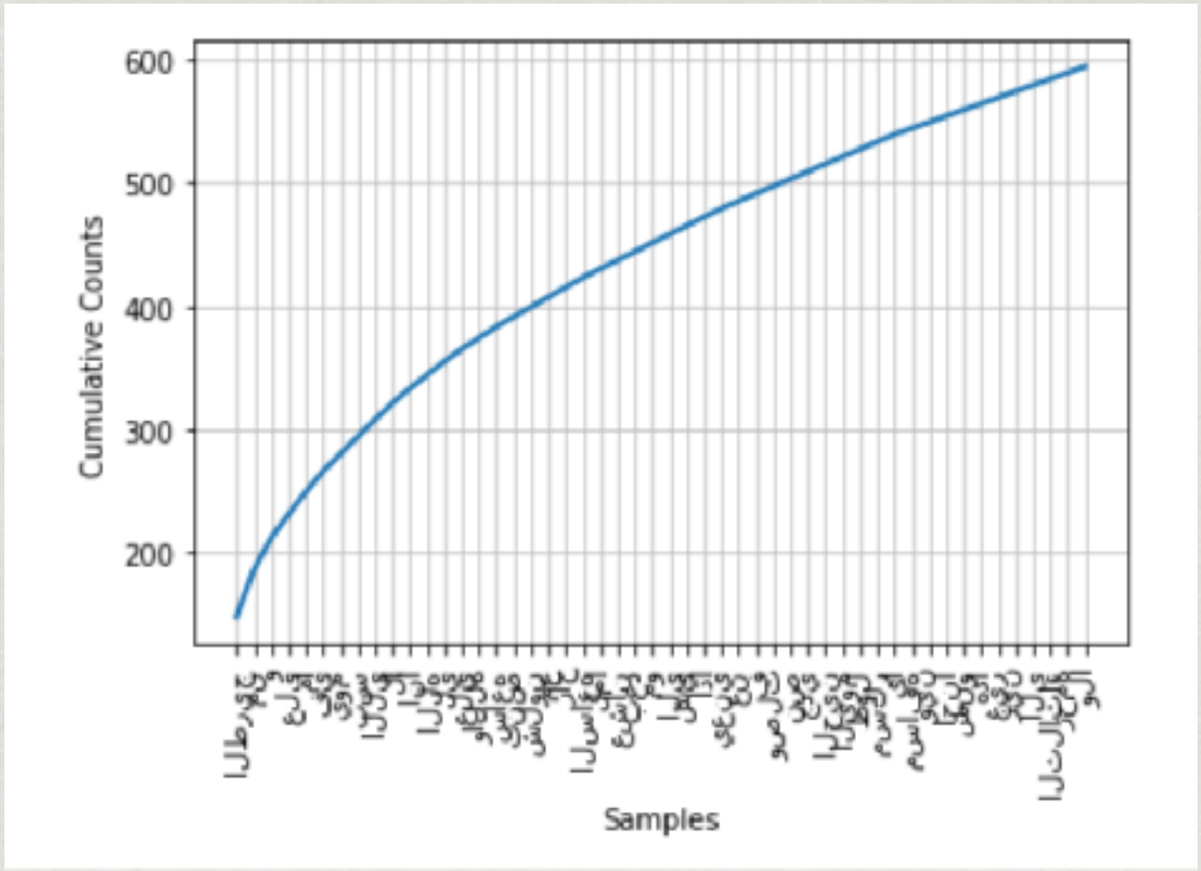
TEXT ANALYSIS & VISUALIZATION

- Corpus size: 300k annotated tweets (50/50 split between EG/GULF).
- Vocabulary size: 356,726



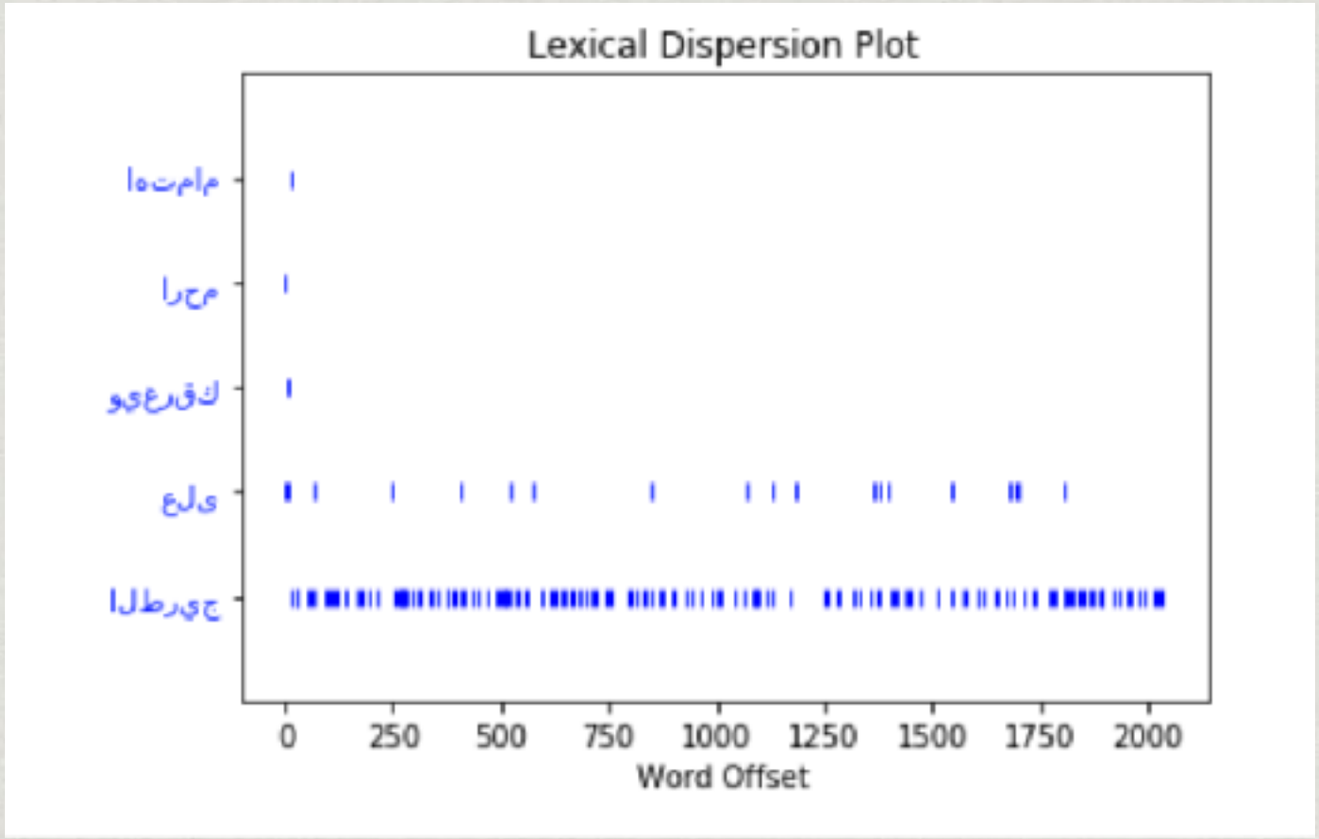
TEXT ANALYSIS & VISUALIZATION

Frequent words can be used to query the streamer with popular dialectal terms not included in the initial keywords list and gather more tweets. Alternatively, they can be added to a stop words list later on.



TEXT ANALYSIS & VISUALIZATION

Check for EG words in GULF subcorpus and vice-a-versa by exploring them in a Lexical Dispersion plot. For example, low word offset for EG terms in GULF subcorpus is a way of inspecting whether or not the classes are sufficiently separated.



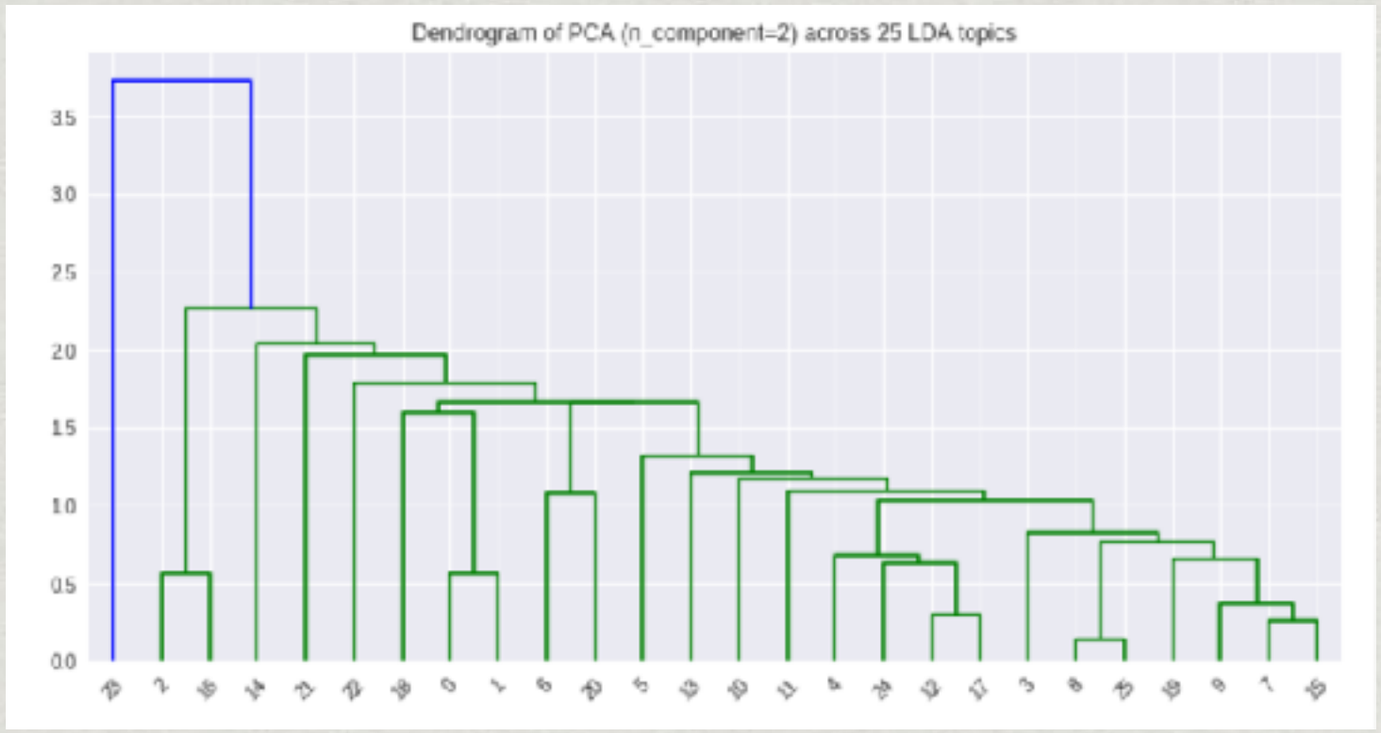
TOPIC MODELING & TRANSFORMATIONS

- Check for term cooccurrences in EG and Gulf documents and add to stop words list.
- Identify dialectically unique keywords and include in the twitter streaming pipeline.
- Test topic models with and without different stop word lists.
- Test topic models with different ranking functions.
- Test pipelines with different classifiers on chosen models.
- Continue rinsing and repeating process as more data comes into the pipeline.

LATENT DIRICHLET ALLOCATION (LDA)

- LDAvis (50 topics) available here: http://34.211.240.206:8888/view/capstone-52/Pickled_from_mongo/lda_75_lem_5_pass.html

```
Top 10 terms for topic #0: ,قمه ,الهلال ,لين ,جمعه ,نادي ,حل ,شر ,الروح ,ماهي
Top 10 terms for topic #1: عبدالله ,وسم ,العربي ,الكثير ,قبي ,القطريين ,اسمع ,انج ,علاقة ,بمناسبة
Top 10 terms for topic #2: يارب ,صالح ,وليس ,آل ,تركي ,الأمير ,شاهد ,يسم ,ملك ,عيوني
Top 10 terms for topic #3: رصيدك ,وانا ,أجمل ,بنت ,قلب ,القطري ,القلوب ,ألف ,السلطان ,العام
Top 10 terms for topic #4: لله ,الحمد ,بخير ,ولم ,البحرين ,القرار ,راسي ,السبب ,دايم ,الأمل
Top 10 terms for topic #5: وما ,مره ,الخليج ,ابن ,تراه ,جمال ,في ,سنه ,عز ,الجو
Top 10 terms for topic #6: يوم ,الحين ,• ,ولو ,المجيد ,🤣🤣🤣 ,هلا ,خالد ,تعال ,جعل
Top 10 terms for topic #7: القلب ,صح ,عادي ,وإن ,الرجال ,المرأة ,الرجل ,صدق ,جمهور ,صوت
Top 10 terms for topic #8: ...يوش ,أفضل ,يسعد ,موقف ,ولن ,يسعدلي ,جالس ,عنا ,مثلك ,ت
Top 10 terms for topic #9: |, تويتر ,فلسطين ,ادري ,يَ ,اهل ,داخل ,الصورة ,سبب ,النفس
Top 10 terms for topic #10: توقعبر ,واجد ,صحيح ,رمي ,احلى ,شعور ,الشخص ,النوم ,يقولون ,بيض
```



MULTINOMIAL NAIVE BAYES

75,735 Test Sample

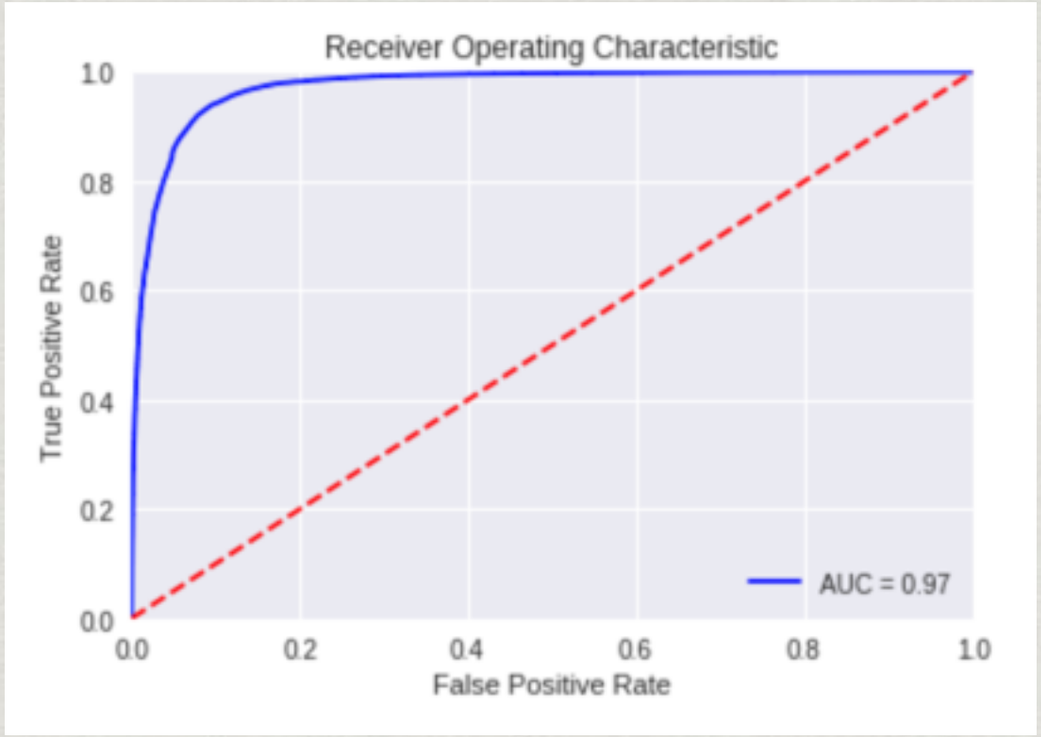
Predicted	0	1	All
True			
0	34328	3631	37959
1	2293	35483	37776
All	36621	39114	75735

92% Prediction Score

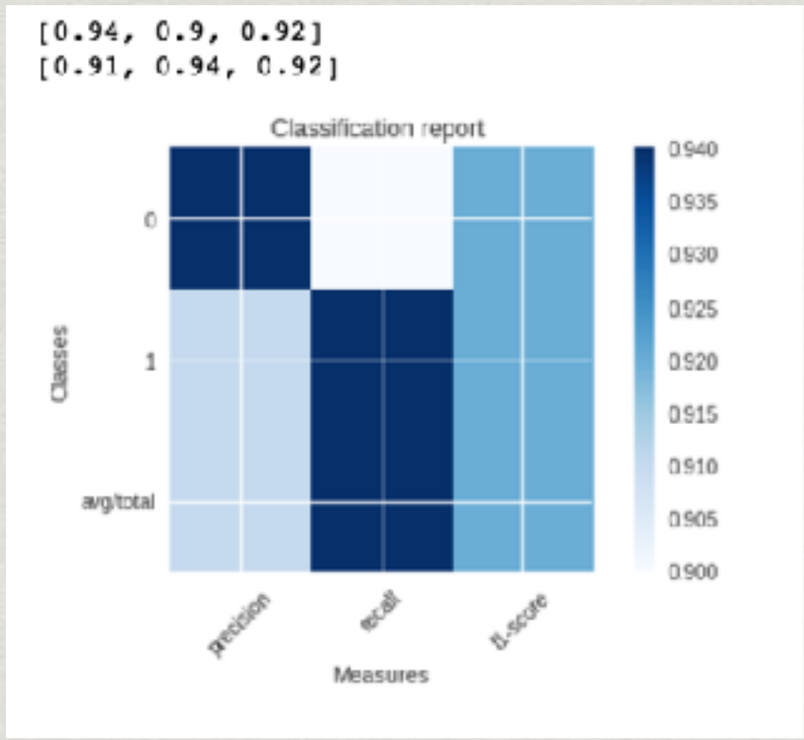
Predicted	0	1	All
True			
0	46.87	4.64	25.06
1	3.13	45.36	24.94
All	50	50	50

MULTINOMIAL NAIVE BAYES

AUC SCORE: 0.97



F1-SCORE: 0.92



NEXT STEPS

- “reflect on how the harvesting process is shaping the later prediction tasks. For example, if a few keywords are driving the decisions about how to annotate users by dialect, then LDA and the classifier are both going to pick up on this. Could this be making the problem artificially easy? Conversely, is it guaranteeing that some classes of mistakes will be made? These are the questions that will propel development of these datasets forward, I think.”
REVIEWER #2 - SRW 2018 Submission

THANK YOU!

TAMIR ELSAHY
GITHUB: TELSAHY
TAMIRELSAHY@GMAIL.COM