Development of Arabic Classifiers for Countering Violent Extremism (Preliminary Research Design Plan)

By Karoline Pershell and Steven Cracknell

*For the purposes of this Research Design Plan, the term Social Media Arabic (**SMA**), refers to the way in which spoken Arabic dialects are typed using the full keyboard in a social media rendering of words, which may include letters, numbers, and punctuation.*

Difficulty with Arabic and Machine Learning.

SMA is the uniting language and the Arabic script is the uniting "script," but as people move to social media and look to quickly type, they are moving to a social media arabic (or SMA) and the SMA is pronunciation based, and therefore has multiple spellings within a dialect, much less across dialects.

Example: baa, which can be spelled in different ways as ba2a, b2a, b'a. This word indicates Egyptian, so may not even be used in other dialects.

The difficulties with Arabic and machine learning are surmountable, but need to be enumerated.

- The same word used in different dialects can may have different meanings. Therefore, we anticipate that it will be necessary to create and train classifiers by dialect, so as to identify words in context.
- The same word, within the same dialect, can have many variations on spelling. In English, we can use roots, stems and a few standard spelling changes (changing a y to an i, or vice versa). These are not yet spelled out for SMA, but perhaps with machine learning we can categorize these spelling variations. For example, baa is also written as b2a, indicating the "two a's". At other times, the number represents a character or a sound (like in English how gr8 is "great"), so this rule is not consistent. (That is, numbers don't only mean to repeat a letter that many times.)

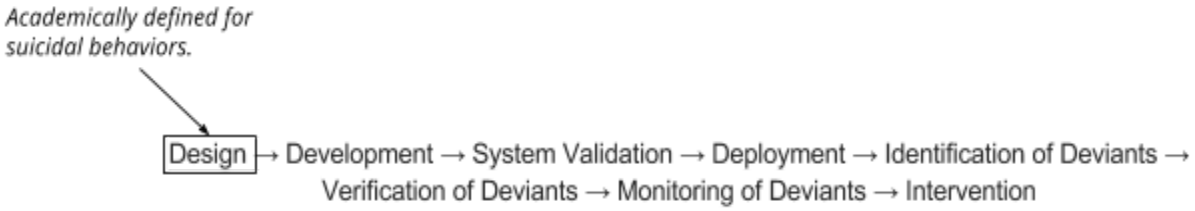
Zenti System Overview

Combining human contextual pattern recognition with machine intelligence, the Zenti system is able to categorize information into subject matter classes, understand communication in context and identify human emotion and intent. The Zenti system is designed for real-time and archived data processing with a highly scalable architecture to process any volume of data, in any language. Using a unique tokenization methodology to weight, score and classify each item of text, the Zenti system calculates 'how much' the underlying text relates to a predefined classifier of interest. The scores are measured in decibels and the more the text is about a class, the 'louder' the result.

The Zenti system is well positioned to process Social Media Arabic, and we already have an extensive development and testing process underway:

Design → Development → System Validation → Deployment → Identification of Deviants → Verification of Deviants → Monitoring of Deviants → Intervention

(1) We have very successfully used this process for identifying deviants, specifically in identification of “people demonstrating suicidal behaviors”.
(2) This has been successful with suicidal behaviors because there is an extensive amount of research around suicide and well-defined risk factors which provided concrete language objectives, making the Design phase very focused.
(3) The identification of CVE follows a nearly identical process to the solution we developed for identifying and targeting Suicidal behaviors.



(4) The primary difference for CVE will be in the Design phase, namely in the type of language and behaviors we are looking for online. To explain the process for CVE, we have used the Suicide behaviour work we’ve done to provide context.
(5) Whilst the process will be described in English, our work in Arabic can be viewed in the appendix.

Walk through of Zenti Process in the context of Suicide

Design. Working closely with Dr. Joe Franklin, his team was able to state the varying language constructs that are red flags for suicidal behaviors.

- While we did have success in defining large classes to identify suicidal behavior (like the early classes based on suicide notes), Dr. Franklin wanted more narrow and subtle classifiers to allow him to test contributions of various factors.
- For example, how someone talks about “Hopelessness” is different than how they talk about “Suicide Ideation,” and by separating those two ideas, he will be able to determine levels of how At-Risk someone is.
- Another benefit to creating narrowly defined classifiers along constructs, is that it eliminates more incorrect samples, that is people who may have said a phrase that is a mash-up of several constructs but is actually meaningless in our context.

- *For CVE*: We could build general classes (e.g., suicide notes of extremists). We may also decide to finely partition the types of language used (e.g., hate speech, logistics planning).

Development. The process of creating classifiers is done through the Zenti Training tool, which is a quickly learned app, developed for subject matter experts (SME's), who don't have to be programmers.

- We have developed a best practices process to easily guide SME's through the Zenti Training process so that they are efficiently developing classifiers and getting feedback on the performance.
- *For CVE:* All of the training will be done in Arabic, by an Arabic language speaker.

System Validation. The Zenti system automatically tests the classifier for accuracy, precision and robustness. This is based on back-end algorithm tokenization and k-fold validation. The system validation phase is used to determine the performance of the classifier as it is trained by the SME. If we find that there are problems with a classifier's performance, we re-engage on the content (typically on the scope of it being too broad or too narrow).

Deployment. Once a classifier has been sufficiently trained, it is deployed to process the live Twitter feed (or any other feed), where the SME can verify and check the incoming data via the Zenti monitoring User Interface.

Identification of Deviants. An SME will review samples that the algorithm finds and flag those that should have further review.

Verification of Deviants. For those individuals who are flagged, the SME can retrieve and score the person's history against a constellation of behavioural classifiers to determine this is an isolated tweet, or if this is a pattern.

- This process can also be automated, so that any single tweet that is flagged, the user history is automatically pulled back and scored.

Monitoring of Deviants. Using a watch list and aggregate behavioural score, SME's can continue to monitor individuals who have high-scoring Tweet histories.

- Depending on SME requirements and use case, an analytical dashboard could be created for long term monitoring (any changes to a person's status, including location, recent posts, changes to profile, etc.).

Intervention. Depending on the severity of the identified user, there are a number of options that can be used for intervention:
To date, We have developed dashboards for semi-automating the intervention process.

- For other clients, intervention has been a the ability to create and send a Tweet with text, an image, and a URL.

- The text has several options pre-written, but any can be edited, or a new one fully re-written on the spot.
- The images typically have a set of several preloaded to choose from (or you can choose to omit an image).
- The URL's take people to a game that is intended to help change their view of self-image. The URL's can be re-routed so that we can track when they have been clicked.

- *For CVE:* Intervention may not be online, but may be turning an individual over to law enforcement.
- In the case of deradicalization, we envision combining the methodologies used for psychological profiling of deviant human behaviour with modern day techniques for social media marketing and targeting.

A few key points to note regarding the development and ownership:

- The Zenti system is a licence technology that is easy to setup and use, and can run on its own independent stack for those who would like to own and use the data.
- We provide the tools and SME training on how to develop the classifiers. This enables the SME to analyze and process data as their need develops.
- Data can be reviewed, maintained, and stored outside of Zenti. That is, **we provide the tools**, but do not need to see the data after collection, nor keep our own records of that.
- Methods of intervention are specific to the context, but if it is **online intervention** (e.g., reaching out to those identified with a text or advertisement), then Zenti can develop tools to semi-automate or automate that process.

How This Will Be Different for CVE

Deviant Identification has been successful for identifying *suicidal behaviors* because the academic literature on the subject carefully spells out the constructs and even language patterns associated with suicide in the US.

That is, if we know what deviants are talking about, then we can find them.

Because the literature on CVE seems less mature than in suicide, we would propose building several classifiers, and then reviewing each to determine whether it correlates to deviant behavior.

Work in Arabic to Date

Zenti's system is based on tokenization that is built from characters, words, and the order of words and groups of words in relation to one another. That is, the system doesn't care whether the characters are Roman, Chinese characters, or Arabic script. A classifier can be trained to identify any combination, so languages can be mixed as well, like French and Arabic, and distinctions can be trained between Modern Standard Arabic and Gulf Arabic.

Zenti created two classifiers in Arabic to identify:

(1) financial/economic headlines, and

(2) headlines announcing disaster (natural or manmade).

Both of these classifiers were trained using mainstream Arabic news providers publishing their headlines to Twitter. The results in the Zenti system show a consistent level of accuracy for the two classifiers.

Arabic Finance

Latest	Analyse		Subject	Unique Count	Score
Jul 26 13:04	ARB Finance		@tdgulf RT @Tdawil_: #موبايلي تسجل صافي ربح بـ 18.8 مليون ريال خلال الربع الثاني #السعودية #تاسي #تداول #الاقتصاد	1	5
Jul 21 17:20	ARB Finance		@nfidk82 11.6 3.2 % مليار ريال أرباح المصارف السعودية في الربع الثاني بتراجع #الاسهم_السعودية #تداول #تاسي	1	3
Jul 21 05:36	ARB Finance		@nfidk53 ارتفاع أرباح "بدجت السعودية" إلى 46.7 مليون خلال الربع الثاني بنسبة 2.3% #تداول #تاسي #الاسهم	1	5
Jul 19 22:44	ARB Finance		@anniebeaurepai1 RT @ma_alhadith: مبيعات العقار تنخفض 29% في الربع الثالث من العام الحالي، بقيمة 80 مليار ريال	3	3
Jul 19 05:39	ARB Finance		@maaalnews 0.1 %نمو أرباح #البنك_السعودي_الهولندي إلى 539.7 مليون خلال الربع الثاني بنسبة #تداول #تاسي #السعودية	1	4
Jul 10 17:25	ARB Finance		@nfidk76 #sabic #tasi #tadawul سابك #السوق_السعودي #سوق_الأسهم #تداول #تاسي #أويشن #توصيات #أسواق #أسهم #تاسي_تداول#	12	2
Jul 07 01:32	ARB Finance		@news_saudis سامسونج تحقق أرباح تشغيل بقيمة 6.98 مليار دولار خلال الربع الثاني من العام الحالي# #السعودية #ksa #saudi	1	2
Jun 29 08:30	ARB Finance		@dailyfx_arabic أبرز مهام مجلس الإحتياطي الفدرالي في الإقتصاد الأمريكي تداول #بورصة #فوركس #دولار #يورو	1	2
Jun 27 22:44	ARB Finance		@ijitha محللون: تراجعات البورصات العالمية تدفع أسواق الخليج للتذبذب #توصيات #تداول #أسهم	1	2
Jun 27 01:31	ARB Finance		@ijitha ساما" تتوقع تباطؤ نمو الناتج المحلي غير النفطي في 2016 #توصيات #تداول #أسهم"	1	3
Jun 22 06:06	ARB Finance		@ijitha إعلان من هيئة السوق المالية بشأن الموافقة على طرح وحدات صندوق استثماري طرحاً عاماً #توصيات #تداول #أسهم	1	3
Jun 21 13:39	ARB Finance		@ijitha مؤشرات الأسهم الأمريكية تغلق على ارتفاع جماعي #توصيات #تداول #أسهم	1	2

Arabic Disaster

Latest	Analyse		Subject	Unique Count	Score
Jul 26 14:57	ARB Disaster		@jagatsinghpur2 الرياض #جدة #مكة #المدينة #ابها #saudi #السعودية #ksa# ملفات مع المذيع محمد الشهري يكشف حقيقة خلافه مع ام بي سي	54	2
Jul 26 14:57	ARB Disaster		@syria_feed #Syria_feed# المرصد_السوري: 11 مقاتلاً على الأقل قضوا في هجوم قوات النظام بالقصف ريف حماة الجنوبي	1	7
Jul 26 14:50	ARB Disaster		@zasffes #وكالة_أعماق مقتل عنصرين من الوحدات الكردية وجرح ثالث بتفجير عبوة ناسفة · #الشدادي #الهلال_يحاجه_قلب_دفاع_اجنبي	1	2
Jul 26 14:49	ARB Disaster		@fgdjkhg5 إسرائيل: غارات استهدفت 4 مواقع لـ #حماس في #غزة ردا على قذيفة سقطت <التفاصيل من هنا ; SYf	1	2
Jul 26 14:40	ARB Disaster		@hamadelhmodi90 الرياض #جدة #المدينة #saudi #السعودية #ksa# رجال الأمن في الحرمين يبذلون جهودا عظيمة في تأمين المعتمرين #مكة ...	11	2
Jul 26 13:48	ARB Disaster		@fawwazmeshefbnv الرياض #جدة #saudi #السعودية #ksa# ... فيديو قناة 24 القضائية مع الحملة الأمنية للقبض على المخمّين والدرباوية بالرياض	2	2
Jul 26 13:42	ARB Disaster		@mak_all RT @Wesal_TV: شاهد كيف تتساقط براميل النظام التصيري على مساكن المدنيين في مدينة #داريا المحاصرة# #سوريا #ريف_دمشق	1	15
Jul 26 13:42	ARB Disaster		@vndgm RT @VideozXtreme: ٧ قتلى و جرحى في #درعا في قصف للطيران السوري	1	2
Jul 26 13:08	ARB Disaster		@hguihhkj البنتاغون: مصرع نائب وزير ورئيس شرطة لدى "داعش" في ضربة جوية قرب الموصل <التفاصيل من هنا ; nHD	1	2
Jul 26 13:02	ARB Disaster		@byj82380 مقتل 14 وأسر جندي من الجيش النيجيري خلال تصدي مقاتلي #الدولة_الإسلامية لهجوم على قرية انقونقون بمنطقة كاجا في ولاية #بورنو شمال شرق #نيجيريا	2	7
Jul 26 12:54	ARB Disaster		@3ajlnet #yemen# عاجل_نت #اخبار #اليوم #اليمن #تعز #عدن #انفجار بمسجد في صنعاء يوقع قتلى وجرحى	1	2
Jul 26 12:47	ARB Disaster		@razuqalmateri65 الرياض #جدة #مكة #المدينة #ابها #الدمام #saudi #السعودية #ksa# برنامج ضيف ون تو تقديم الغانم و العقيلي مع المصيبيح	21	2
Jul 26 12:42	ARB Disaster		@qasiounnewsar وكالة_قاسيون	#اخر_الأخبار #اخر الأخبار: وكالة أعماق: تنظيم الدولة الإسلامية يستعيد السيطرة على قرية اليالنجي بعد اشتباكات مع...	1	10

Validation and verification regarding identification of Arabic sentiment is currently underway. A review of this work will be included as an appendix to this summary.

Planned Future Classifier Development and Testing

We would be interested in working with SME's to further develop this list, but a first attempt at behaviors/language of interest is below:

- Suicide ideation
- Violence Ideation
- Religious Proselytizing
- Hate speech
- Event Preparation (for this, we would begin with the active shooter categories for language, see the example on Knowledge Mapping. How would this be adjusted for CVE?)

Previous claims about natural language processing in Arabic are overstated. To demonstrate that Zenti understands this and is creating a more accurate machine, we propose development of classifiers and testing of classifiers as detailed in this separate document.