Supporting Civilian Protection: A Machine Learning System for Detecting Evidence of Airstrikes on Social Media

Rickard Nyman, Franz Busse, David Levin, Daniel Henebery
Hala Systems Inc

Abstract

This project describes work carried out to support a civilian protection system called Sentry, which generates warnings in advance of impending airstrikes in Syria. Critical to the system is information gathering with high precision and high recall. Evidence of incidents posted on social media are monitored and vetted by a group of human experts. To improve the process and increase scalability, Hala Systems has built and deployed a machine learning-based system that automatically ingests data from Facebook, Telegram and Twitter and detects and aggregates evidence of airstrikes in Syria. The detection model achieves an out-of-sample F1 score of 0.96 leading to a reduction in the average time spent by human researchers by 50-80%, thus increasing the scalability and accuracy of the system as a whole.

Introduction

Hala Systems develops solutions to protect civilians and provide accountability in conflict zones. Through the development and implementation of innovative technology, Hala aims to reduce harm, increase security, combat disinformation, and stabilize communities. Hala's early warning system in Syria, called Sentry, is used to observe hostile aircraft, analyze their attack patterns, and send warnings to civilians of imminent attack. Over the past four years, this system has saved hundreds of civilian lives, and amassed critical information used by the UN and other organizations that seek accountability in volatile situations [1]. According to a preliminary analysis, Hala's technology solutions resulted in an estimated reduction of 20 to 30% in total casualties in several areas under heavy bombardment in 2018 and 2019 [2].

Social media posts are an important input for the system. The system is constantly ingesting posts regarding airstrikes in the conflict region, and then correlates those reported airstrikes with observed aircraft activity. This serves two purposes: first, it helps train the machine learning prediction algorithms so that the system can provide more accurate warnings of future strikes; second, it is a crucial element for bringing accountability for strikes against civilian targets. Originally, this process was performed by human monitors who spent a significant amount of time searching online media for information about airstrikes. However, this is very time consuming, and the volume is large enough that the human monitors may not be able to read everything. This is costly, unscalable, and less accurate. For these reasons, Hala needed a capability to automate the search and identification of open media reports of airstrikes in Syria.

At the time this project began (and to the best of our knowledge this remains the case), popular existing tools for carrying out text analysis and natural language processing (NLP) rarely supported Arabic, with some notable exceptions. For example, the now highly cited Stanford project *Global Vectors for Word Representation* (GloVe) [3], provides downloadable databases of word vector embeddings only in English. The databases are derived from co-occurrence analysis of billions of terms from different corpora, including Twitter and Wikipedia. The success of these methods and other recent influential developments in NLP for deriving word vector embeddings using neural networks [4] motivated a project to derive similar databases of vector embeddings in Arabic, named AraVec [5].

[6] Alabbas et al. carry out a systematic review of Arabic text classification, highlighting many of the idiosyncratic challenges particular to this language. Models typically found in the literature include support vector machines (SVM), k-NN, naïve Bayes and decision trees, with SVM typically found to perform the best with accuracy scores between 80-95%. Common datasets used in these studies included the Quran, other religious scripts and old books, websites and news articles. Not many studies have attempted to perform text classification on Arabic social media data, with the notable exceptions of [7] and [8] on sentiment and dialect classification, respectively.

This motivated us to develop our own automated system for scanning through Arabic content of open media platforms and flagging all posts that mention airstrikes in Syria.

Methodology

Overview

We develop a pipeline machine learning model aimed at detecting social media posts in Arabic mentioning airstrikes in Syria. We assess the ability of various machine learning models and feature representations, including the use of word vector embeddings derived for the Arabic language, like those derived in [5].

Data for the task were collected from a selection of 127 public Facebook pages. The specific set of sources was initially selected by human experts and then expanded on by automatically detecting references to additional Facebook pages within a sample of the scraped data. Using the Facebook Graph API *https://developers.facebook.com/docs/graph-api/* we were able to scrape each Facebook page back to its first published post. Following this procedure, we were able to gather 1,513,000 posts with the earliest one posted in March, 2011. See Figure 1a for the total number of posts per day.

This paper is focused on an analysis of a subset of the collected Facebook posts that were categorized by experts as either mentioning an airstrike in Syria or not. The posts are pre-processed using the method in [5] before being transformed into the various feature representations and finally passed to a machine learning classifier.

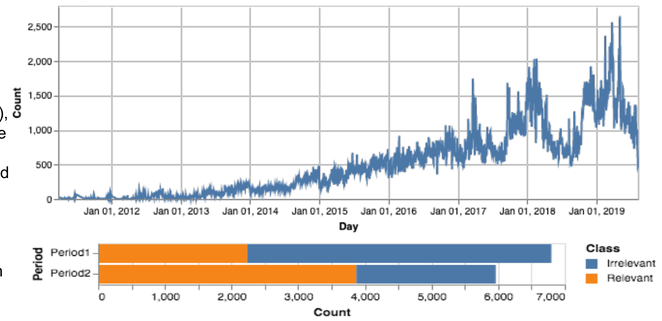
Data

A set of 12,763 Facebook posts were labelled by human experts as part of their daily routine: 6111 relevant (about airstrikes) posts and 6652 irrelevant posts (anything else on the Facebook page).

The labelled data selected for the task were collected over two distinct periods. Due to content drift, many machine learning models quickly deteriorate over time. Using data from two periods separated in time allows us to test the ability of different models to generalise over longer periods, therefore ensuring the selected model relies on more time-invariant features when making its classifications.

Figure 1b shows the sample breakdown: 6799 posts collected from September to October 2017 (period 1): 2239 relevant and 4560 irrelevant; 5964 posts collected from March to June, 2018 (period 2): 3872 relevant and 2092 irrelevant.

Figure 1



Models and feature representations

We compare a range of feature representations and machine learning models, including methods typically found to be suitable for text classification tasks in Arabic, but also some which are less commonly found in the literature to date.

Feature representations considered include common unigram term-weighting schemes: boolean, term frequency (tf) and term frequency - inverse document frequency (tf-idf); topic/semantic model representations: latent Dirichlet allocation and latent semantic analysis; and word embeddings from our own word2vec model trained on the full set of Facebook posts in the database as well as those provided by the authors of AraVec (*https://github.com/bakrianoo/aravec*).

When constructing feature vectors with different term weighting schemes the terms were selected both by using a Chi-square test as well as expert judgement. In both cases, approximately 200 relevant terms were selected.

To construct feature vectors using word embeddings, we average the term vector embeddings for all the words in the post. Several different dimensions of the embeddings were considered, and in the case of word2vec, we learnt embeddings using both continuous skip-gram and continuous bag-of-words, as well as by experimenting with different lengths of the context word window.

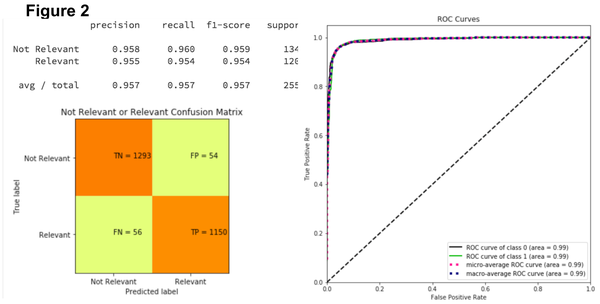
A broad range of machine learning algorithms were considered, including naïve Bayes (Gaussian, Multinomial and Bernoulli), logistic regression, random forest, gradient boosted trees and linear and RBF support vector machines. All models were tuned by searching over a grid of parameter combinations.

Evaluation

We are interested in measuring both precision and recall of the system. Precision is measured as the proportion of posts correctly detected by the system as relevant and recall is measured as the proportion of all relevant posts detected by the system. We evaluate the success of a model using the standard F1 metric defined as the harmonic mean of precision and recall, F1 = 2/[1/recall+1/precision].

We set aside a random subsample consisting of 20% of the data for final validation of the best performing model. Using the remaining 80% of observations we evaluate each model and feature combination on two different criteria: • we measure the mean and standard deviation F1 using 10-fold cross-validation (*out-of-sample*) • due the highly dynamic nature of our problem (a rapidly changing environment, .e.g., changing targets, locations, aircrafts, etc.) we also measure the F1 score of each model trained on data from one period (Period 1) and evaluated on data from the other period (Period 2), and vice versa ('out-of-period'). The latter is a critical test.

Figure 2



Results

The mean F1 10-fold cross validation scores ranged from 0.60 up to 0.96 for all combinations of features and models. All models received higher F1 scores when trained on data from the first period and evaluated on the second period, compared to the scores obtained when training on data from the second period and evaluated on data from the earlier period. All models can be seen to deteriorate over time, but the best performing model still achieved an out-of-period average F1 score of 0.91 (i.e., when averaging the scores of using both periods as the evaluation set).

In general, the feature representations derived from word2vec term embeddings perform well across models. The representations derived from AraVec's term embeddings also do well, despite not being derived from this particular corpus. Unigram vector representations with boolean term weights tend to do well, consistent with results of text classification in many languages, including Arabic.

Among the different machine learning models, the linear SVM consistently achieves high results, in line with other results for text classification in Arabic. Both the random forest and gradient boosted trees also achieve state-of-the-art results, indicating that tree-based ensemble models are also suitable candidates for Arabic text classification tasks.

We finally select a model that performed well across both evaluation criteria, achieving both high precision and high recall on the pooled data, as well as exhibiting the least drift in performance over time. Figure 2a shows the confusion matrix with the associated performance metrics when applying the final model on the 20% held out dataset, i.e., the dataset set aside at the start of the project and therefore not seen by the model beforehand. Figure 2b shows the corresponding ROC curve.

Conclusion

The success of the Sentry early warning system depends on accurate and complete information on airstrikes in Syria.

In this work, we present the results of a machine learning system to automate the task of gathering such data. As a result of testing a wide range of feature representations and machine learning classifiers the best performing model achieves an out-of-sample F1 score of 0.96, putting it in line with the top performing Arabic text classifiers.

Feedback from researchers who directly use the output of this system indicates it has reduced the time and effort needed to investigate an airstrike by reducing the set of posts which need to be examined, automatically extracting key features, and consolidating disparate sources in one place. The lead of the research team estimates that the model has reduced the time it takes to investigate an airstrike by 50 to 80%. A secondary benefit is to the mental health of the research team. Frequently-cited sources often contain macabre imagery depicting the aftermath of airstrikes. The output of the model necessarily acts as a layer of abstraction, thereby reducing the researchers exposure to these images.

The applicability of a system like the one presented here is not specific to Sentry or Syria. Event datasets derived from social media are increasingly common and widely used. The development of these datasets can and should be improved using advances in machine learning and artificial intelligence as this information helps the most vulnerable in society by detecting violent acts and providing accountability for them.