Ahmad Mustafa

Jordan
(+962) 79-095-3097 • ahmad.mo.mustafa@gmail.com

Education

- **The University of Texas at Dallas** — **Richardson, TX**
  PhD in Computer Science — *May 2018*
- **The University of Texas at Dallas** — **Richardson, TX**
  Master in Computer Science — *2015*
- **Jordan University of Science and Technology** — **Irbid, Jordan**
  Master in Computer Sciencee — *2010*
- **Yarmouk University** — **Irbid, Jordan**
  Bachelor in Computer Information Systems — *2007*

Experience

- **Jordan University of Science and Technology** — **Irbid, Jordan**
  Assistant Professor — *Sep 2019 – Present*
- **Mawdoo3.com** — **Amman, Jordan**
  AI Research Scientist — *May 2018 – Aug 2019*
- **The University of Texas at Dallas** — **Richardson, TX**
  Research Assistant — *2012 – May 2018*
  Duties include working on several Machine Learning projects in multiple domains; including data streams, natural language processing, text mining, image/video analytics, cyber security, and network traffic. **Details**:
  - Cross-lingual Duplicate and Near-Duplicate Detection in Textual News Reports using Multilingual Word Embedding (Bilingual FastText), NLP techniques, Clustering (DBSCAN and Kmeans), Text similarity, Outlier detection techniques, and Supervised and Unsupervised classification. [Python]
  - Working with infinite flow of data points (data streams) like Textual News reports, web-traffic data, and network-based system attacks.
  - Detecting new (Novel) classes that have not been previously modeled by classifiers in Evolving Data Stream environment. [Java]
  - Using Statistical Change point detection methods for Classification and updating classification models. [Java]
  - Encrypted network traffic characterization and fingerprinting. [Python]
  - HostWatch: Situational Awareness of Machine State for Cybersecurity: Developed Anomaly Detection methods using Bayesian and Markov Networks to detect network attacks by analyzing System calls collected from victim's machine. [Java]
- **Intel Corp.** — **Hillsboro, OR**
  Data Scientist Intern — *May 2014 – Jan 2015*
  - Workload Characterization: Analyzing software, hardware, and network performance data using machine learning techniques
  - Data collection of CPU, memory, and network data using EMON and SAR tools while running benchmark workloads for a certain time period. [Python]
  - Data cleaning of CPU, memory, and network data by aligning attributes of data points according to their time-stamp and applying interpolation. [R Scripting language]
  - Analyzing the performance of CPU, memory, and network data using clustering, linear regression, and several other machine learning techniques. [R Scripting language]
  - Studying the statistical correlation between CPU, memory, and network variables. [R Scripting language]
  - Analyzing the impact of Java variables on CPU utilization. [R Scripting language]
  - Tracking software phases by monitoring CPU utilization using Kmeans clustering [R Scripting language]
  - Analyzing the performance of Ensemble-based classification models like Random Forest on multi-core CPU. [R Scripting language]
  - Using anomaly detection techniques to detect scam phone calls. [R Scripting language]

- **Jordan University of Science and Technology** — **Irbid, Jordan**
  Teaching Assistant — *Sep 2009 – Jan 2010*

Research Interests

Artificial Intelligence, Data Mining, Natural Language Processing (NLP), Deep learning, Data Stream Mining, Outlier Analysis, Cyber Security.

Publications

- **Cross-Lingual Duplicate Detection Over News Reports**
  work-in-progress
- **Automating Cyberdeception Evaluation with Deep Learning**
  To appear in Proc. 53rd Hawaii International Conference on System Sciences (HICSS) January 2020
  A machine learning-based methodology is proposed and implemented for conducting evaluations of cyberdeceptive defenses with minimal human involvement. This avoids impediments associated with deceptive research on humans, maximizing the efficacy of automated evaluation before human subjects research must be undertaken. Leveraging recent advances in deep learning, the approach synthesizes realistic, interactive, and adaptive traffic for consumption by target web services. A case study applies the approach to evaluate an intrusion detection system equipped with application-layer embedded deceptive responses to attacks. Results demonstrate that synthesizing adaptive web traffic laced with evasive attacks powered by ensemble learning, online adaptive metric learning, and novel class detection to simulate skillful adversaries constitutes a challenging and aggressive test of cyberdeceptive defenses.
- **Deep Contextualized Pairwise Semantic Similarity for Arabic Language Questions**
  Proceedings of the 31st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2019)
  https://arxiv.org/abs/1909.09490
  Question semantic similarity is a challenging and active research problem that is very useful in many NLP applications, such as detecting duplicate questions in community question answering platforms such as Quora. Arabic is considered to be an under-resourced language, has many dialects, and rich in morphology. Combined together, these challenges make identifying semantically similar questions in Arabic even more difficult. In this paper, we introduce a novel approach to tackle this problem, and test it on two benchmarks; one for Modern Standard Arabic (MSA), and another for the 24 major Arabic dialects. We are able to show that our new system outperforms state-of-the-art approaches by achieving 93 % F1-score on the MSA benchmark and 82 % on the dialectical one. This is achieved by utilizing contextualized word representations (ELMo embeddings) trained on a text corpus containing MSA and dialectic sentences. This in combination with a pairwise fine-grained similarity layer, helps our question-to-question similarity model to generalize predictions on different dialects while being trained only on question-to-question MSA data.
- **NSURL-2019 Shared Task 8: Semantic Question Similarity in Arabic**
  Proceedings of the First Workshop on NLP Solutions for Under Resourced Languages, NSURL '19 Trento, Italy (2019).
  https://arxiv.org/abs/1909.09691
  Question semantic similarity (Q2Q) is a challenging task that is very useful in many NLP applications, such as detecting duplicate questions and question answering systems. In this paper, we present the results and findings of the shared task (Semantic Question Similarity in Arabic). The task was organized as part of the first workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) The goal of the task is to predict whether two questions are semantically similar or not, even if they are phrased differently. A total of 9 teams participated in the task. The datasets created for this task are made publicly available to support further research on Arabic Q2Q.
- **Unsupervised Deep Embedding for Novel Class Detection over Data Stream**
  2017 IEEE International Conference on Big Data, Big Data 2017, Boston, MA, USA, December 11-14, 2017.
  http://ieeexplore.ieee.org/document/8258127
  Data streams are continuous flows of data points. Novel class detection is an important part of data stream mining. A novel class is a newly emerged class that has not previously been modeled by the classifier over the input stream. This paper proposes deep embedding for novel class detection—a novel approach that combines feature learning using denoising autoencoding with novel class detection. A denoising autoencoder is a neural network with hidden layers aiming to reconstruct the input vector from a corrupted version. A nonparametric

multidimensional change point detection approach is also proposed, to detect concept-drift (the change of data feature values over time). Experiments on several real datasets show that the approach significantly improves the performance of novel class detection.

- **Near real-time atrocity event coding**
  2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 2016, pp. 139-144.
  http://ieeexplore.ieee.org/document/7745457
  In recent years, mass atrocities, terrorism, and political unrest have caused much human suffering. Thousands of innocent lives have been lost to these events. With the help of advanced technologies, we can now dream of a tool that uses machine learning and natural language processing (NLP) techniques to warn of such events. Detecting atrocities demands structured event data that contain metadata, with multiple fields and values (e.g. event date, victim, perpetrator). Traditionally, humans apply common sense and encode events from news stories but this process is slow, expensive, and ambiguous. To accelerate it, we use machine coding to generate an encoded event. In this paper, we develop a near-real-time supervised machine coding technique with an external knowledge base, WordNet, to generate a structured event. We design a Spark-based distributed framework with a web scraper to gather news reports periodically, process, and generate events. We use Spark to reduce the performance bottleneck while processing raw text news using CoreNLP.

- **Adaptive encrypted traffic fingerprinting with bi-directional dependence**
  The 32nd Annual Conference on Computer Security Applications (ACSAC '16). ACM, New York, NY, USA.
  https://dl.acm.org/citation.cfm?id=2991123
  Recently, network traffic analysis has been increasingly used in various applications including security, targeted advertisements, and network management. However, data encryption performed on network traffic poses a challenge to these analysis techniques. In this paper, we present a novel method to extract characteristics from encrypted traffic by utilizing data dependencies that occur over sequential transmissions of network packets. Furthermore, we explore the temporal nature of encrypted traffic and introduce an adaptive model that considers changes in data content over time. We evaluate our analysis on two packet encrypted applications: website fingerprinting and mobile application (app) fingerprinting. Our evaluation shows how the proposed approach outperforms previous works especially in the open-world scenario and when defense mechanisms are considered.

- **Recurring and Novel Class Detection Using Class-Based Ensemble for Evolving Data Stream**
  IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 10, pp. 2752-2764, Oct. 1 2016.
  http://ieeexplore.ieee.org/document/7350165/
  Streaming data is one of the attention receiving sources for concept-evolution studies. When a new class occurs in the data stream it can be considered as a new concept and so the concept-evolution. One attractive problem occurring in the concept-evolution studies is the recurring classes from our previous study. In data streams, a class can disappear and reappear after a while. Existing studies on data stream classification techniques either misclassify the recurring class or falsely identify the recurring classes as novel classes. Because of the misclassification or false novel classification, the error rates increases on those studies. In this paper we address the problem by defining a novel ensemble technique "class-based" ensemble which replaces the traditional "chunk-based" approach in order to detect the recurring classes. We discuss the details of two different approaches in class-based ensemble and explain and compare them in detail. Different than the previous studies in the field, we also prove the superiority of both "class-based" ensemble method over state-of-art techniques via empirical approach on a number of benchmark data sets including Web comments as text mining challenge.

- **Evolving Stream Classification using Change Detection**
  10th IEEE International Conference on Collaborative Computing, Miami, FL, 2014, pp. 154-162.
  http://ieeexplore.ieee.org/document/7014560/
  Classifying instances in evolving data stream is a challenging task because of its properties, e.g., infinite length, concept drift, and concept evolution. Most of the currently available approaches to classify stream data instances divide the stream data into fixed size chunks to fit the data in memory and process the fixed size chunk one after another. However, this may lead to failure of capturing the concept drift immediately. We try to determine the chunk size dynamically by exploiting change point detection (CPD) techniques on stream data. In general, the distribution families before and after the change point are unknown over the stream, therefore non-parametric CPD algorithms are used in this case. We propose a multi-dimensional non-parametric CPD technique to determine chunk boundary over data streams dynamically which leads to better models to classify instances of evolving data streams. Experimental results show that our approach can detect the change points and classify instances of evolving data stream with high accuracy as compared to other baseline approaches.

- **Host-Based Anomaly Detection Using Learning Techniques**
  2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, 2013, pp. 1153-1160.
  http://ieeexplore.ieee.org/document/6754055/
  Anomaly detection is a crucial part of computer-security. This paper presents various host based anomaly detection techniques. One technique uses clustering with markov network (CMN). In CMN we first cluster the benign training data and then from each cluster we build a separate markov network to model the benign behavior. During testing, each Markov network calculates the probability of each testing instance. If the probability from multiple markov networks is low, we classify the point as malicious. The paper also presents CMN with Outlying subspace (CMN-OS). In CMN-OS, a training data set that consists of benign and few malicious data is used to identify the outlying subspace which is used as a lower dimensional representation of the full dimensional space. Then, CMN uses the new subspace to represent its training and testing data sets. Finally, the paper presents Clustered Label Propagation (CLP). CLP starts by clustering benign and malicious training. It then labels each cluster based on its central-most point. During testing, these points are added to the testing data as labeled points and Label Propagation is used to label the testing data. We experimentally show that CMN approach outperforms several other approaches and performs similar to CMN-OS.

- **Novel Class Detection and Feature via a Tiered Ensemble Approach for Stream Mining**
  2012 IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI), Athens, 2012, pp. 1171-1178.
  http://ieeexplore.ieee.org/document/6495184/
  Static data mining assumptions with regard to features and labels often fail the streaming context. Features evolve, concepts drift, and novel classes are introduced. Therefore, any classification algorithm that intends to operate on streaming data must have mechanisms to mitigate the obsolescence of classifiers trained early in the stream. This is typically accomplished by either continually updating a monolithic model, or incrementally updating an ensemble. Traditional static data mining algorithms futile in a streaming context (and often in a distributed sensor network) due to their need to iterate over the entire data set locally. Our approach – named HSMiner (Hierarchical Stream Miner) – takes a hierarchical decomposition approach to the ensemble classifier concept. By breaking the classification problem into tiers, we can better prune the irrelevant features and counter individual classification error through weighted voting and boosting. In addition, the atomic decomposition of feature inputs enables straightforward mapping to distributing the ensemble among resources in the network. The implementation proves to be fast and very memory conservative, and we emulate a distributed environment via signal-linked threads. We examine the theoretical and empirical analysis of our approach, specifically examining trade-offs of three different novel class detection variations, and compare these results to a similar method using benchmark data sets.