Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 275-284 (2017)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning

Batoul Haidar*,[1], Maroun Chamoun[1], Ahmed Serhrouchni[2]

[1]Saint Joseph University, Lebanon

[2]Telecom ParisTech, France

ARTICLEINFO

Article history:
Received: 12 November, 2017
Accepted: 03 December, 2017
Online: 23 December, 2017

Keywords:
Cyberbullying
Machine Learning
Natural Language Processing
Arabic Natural Language Processing

ABSTRACT

With the abundance of Internet and electronic devices bullying has moved its place from schools and backyards into cyberspace; to be now known as Cyberbullying. Cyberbullying is affecting a lot of children around the world, especially Arab countries. Thus, concerns from cyberbullying are rising. A lot of research is ongoing with the purpose of diminishing cyberbullying. The current research efforts are focused around detection and mitigation of cyberbullying. Previously, researches dealt with the psychological effects of cyberbullying on the victim and the predator. A lot of research work proposed solutions for detecting cyberbullying in English language and a few more languages, but none till now covered cyberbullying in Arabic language. Several techniques contribute in cyberbullying detection, mainly Machine Learning (ML) and Natural Language Processing (NLP). This journal extends on a previous paper to elaborate on a solution for detecting and stopping cyberbullying. It first presents a thorough survey for the previous work done in cyberbullying detection. Then a solution that focuses on detecting cyberbullying in Arabic content is displayed and assessed.

1. Introduction

As stated in the Abstract, this work is an extension to the author's previous work [1] which was published in the EMS2016 conference.

Children and teens were exposed to physical bullying before the abundance of internet, computers and handheld devices. Nowadays, bullying is performed using cyber technology. Around 50% of the youth of America are suffering from cyberbullying [2]. As for the Arab world: 20.9% of middle-school adolescents report bullying in UAE, 31.9% in Morocco, 33.6% in Lebanon, 39.1% in Oman and 44.2% in Jordan [3].

Awareness of cyberbullying is rising in the Arab world. Arab News [4] declares that they heard of numerous cyberbullying incidents in Saudi Arabia. One of the rare reports [5] on cyberbullying states that 60% of Gulf Countries' youth openly admit the presence of cyberbullying amongst their peers. This study also states that only quarter the predators online do bully their victims offline. Which means that internet have encouraged three quarters of the predators to bully others, while they wouldn't have considered bullying face to face.

Most of the previous research dealing with cyberbullying focused on the effects of cyberbullying. It was concerned with helping the victims after cyberbullying attacks, mainly from the psychological aspect. Less work was directed towards implementing technical methods to detect and stop an ongoing cyberbullying attack, or even to prevent cyberbullying attacks while or before they happen [6]

After the Introduction, this paper furnishes a background that covers the technologies underlying cyberbullying detection in Section 2. Then in section 3 presents a survey of all existing literature in both cyberbullying detection and multilingual techniques. It concludes at the end of section 3 that there is no work done on detecting cyberbullying attacks in Arabic language. Therefore, in section 4 a system for cyberbullying detection in Arabic content is proposed. This system is designed for preventing cyberbullying attacks, by detecting and stopping them. It uses Natural Language Processing (NLP) to identify and process Arabic words. Then Machine Learning (ML) techniques are used to classify bullying content. In section 5 the results obtained from the system are displayed and analyzed. Conclusions and future works are finally stated in section 6.

2. Background

2.1. Cyberbullying

Cyberbullying is defined as the use of Internet, cell phones, video game systems, or other technologies to send or post text or

* Batoul Haidar: batoul.haidar@net.usj.edu.lb

images intended to hurt or embarrass another person or group of people [7]. Some examples of cyberbullying include sending mean or threatening messages, tricking someone into revealing personal or embarrassing information and sending it to others, sending or forwarding private messages to others, sharing explicit pictures with others without consent, starting rumors via text message or online or creating fake online profiles on websites such as Facebook, Myspace, Twitter, etc. to make fun of people [7]. There are several categories of cyberbullying as stated by [8] and [9] :

- Flaming: starting a form of online fight.
- Masquerade: where there is a bully pretending to be someone else, in order to perform malicious intents.
- Denigration: sending or posting gossip to ruin someone's reputation.
- Impersonation: Pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person's reputation or friendships.
- Harassment: Repeatedly sending profane and cruel messages.
- Outing: Publishing someone's embarrassing information, images or secrets.
- Trickery: Talking someone into revealing secrets or embarrassing information for the sake of sharing it online.
- Exclusion: Intentionally and cruelly excluding someone from an online group.
- Cyberstalking: Repeated, intense harassment and denigration that includes threats or creates significant fear.

Bullying and cyberbullying leave mental and physical effects on both the bully (predator) and the victim. Cyberbullying is more severe than physical bullying due to the fact that it is wider, public, and the victim has nowhere to escape. Victims of cyberbullying reported emotional, concentration, and behavioral issues, as well as trouble getting along with their peers. These victims were more likely to report frequent headaches, recurrent stomach pain, and difficulty sleeping. One out of four students revealed that they felt unsafe at school. They were also more likely to be hyperactive, have conduct problems, abuse alcohol, and smoke cigarettes [10].

As stated previously, cyberbullying detection systems mainly use Machine Learning and Natural Language Processing techniques in the course of detection. Thus before dwelling in the previous literature in the area of cyberbullying detection, a thorough background for ML, NLP and other techniques collaborating in the process of cyberbullying detection is displayed.

2.2. Machine Learning

Machine Learning (ML) is defined as the ability of a computer to teach itself how to take a decision using available data and experiences [11]. Available Data is known as *Training Data*. Decisions to be taken in ML might be a classification or prediction for new objects or data. The computer classifies a new piece of data by depending on learning algorithms. When the training data is labeled, i.e. classified by human experts, the algorithms depending on these labeled data are called Supervised Learning algorithms [12].

In cyberbullying detection, there could be a corpus of data manually labeled (or classified) by people as either containing harm or not, as described in Section 3. When the training data is unlabeled, the algorithms depending on these non-labeled data are called Unsupervised Learning algorithms [12]. They teach themselves how to classify the data based on similarities and differences between data. When both supervised and unsupervised learnings are combined together by using labeled and unlabeled data, to get the most out of both ways, the algorithm is known as Semi-supervised Learning algorithm [12].

When ML is used to classify a certain object as belonging or not belonging to a certain category, the machine learner is called *Binary Classifier* [13]; for example in spam email filtering, ML algorithms are used to take decisions against incoming emails and label them as either spam or not spam. A second type is when the task given to the classifier, is to match a certain object against several classes or categories, then it is called *Multi-Class Classifier*. A third type might be predicting a value for an object and is called *Regression*, i.e. predicting a priority level for an incoming email.

There are several ML algorithms available, from which the most frequently used in relevance to the scope of research of this paper will be mentioned.

- Naive Bayes: A probabilistic supervised learning method [14] that mainly calculates the probability of an item belonging to a certain class, depending on metrics obtained from training data. Naïve Bayes algorithm was used in some cyberbullying detection research, such as in [15] and [14]. It was used for sexual predation detection.
- Nearest Neighbor Estimators: A simple estimator [16] that uses distance between data instances, in order to map a certain instance to its closest distance neighbor, thus estimating the class of this instance, this algorithm was used in [17] and [18].
- Support Vector Machine (SVM): Also a supervised algorithm. SVM is a binary classifier that assumes a clear distinction between data samples. It tries to find an optimal hyper plane that maximizes the margin between the classes [12]. SVM was used in many cyberbullying detection systems [19] , [20].
- Decision Tree: Decision tree learners use a set of labeled data, thus they are supervised learners. Decision trees classify data using a command and conquer approach. The Trees compose of leaves and arcs. Each leaf of the tree represents a classification class and each arc represents a feature inspected from training data [21]. The C4.5 algorithm is an implementation of decision trees. It was employed in cyberbullying detection by [22] and [17].

ML algorithms are widely incorporated in cyberbullying detection systems as seen in Section III, due to the huge amount of

data incorporated in social networking platforms, which makes it hard to be processed by human power, thus comes the need for a machine learner.

2.3. Natural Language Processing

Natural Language Processing (NLP) is the collection of techniques employed to make computers capable of understanding the natural unprocessed language spoken between humans by extracting grammatical structure and meaning from input [23].

NLP is a common branch of all of Linguistics, Artificial Intelligence and Computer Science [24]. NLP research started with Machine Translation in the late 1940s [25]. Then it spread to other areas of application, such as information retrieval, text summarization, question answering, information extraction, topic modeling, opinion mining [26], optical character recognition, finding words boundary, word sense disambiguation, and speech recognition [24].

According to Chandhana [27], NLP can be divided into three areas; Acoustic – Phonetic: where acoustic knowledge studies rhythm and intonation of language; i.e. how to form phonemes, the smallest unit of sounds. Phonemes and phones are aggregated into word sounds. Phonetic knowledge relates sounds to the words we recognize. Morphological – Syntactic: Morphology is lexical knowledge which studies sub words (morphemes) that would form a word. Syntactic knowledge studies the structural roles of words or collection of words to form correct sentences. Semantic - Pragmatic: Semantic knowledge deals with the meaning of words and sentences, while pragmatic knowledge deals with deriving sentence meanings from the outside world or outside the content of the document [28].

2.4. Sentiment Analysis

Sentiment analysis is a textual analysis technique. Sentiment analysis is used to define the polarity, subjectivity or features of a certain text. By polarity we mean defining whether a certain content is positive, negative or neutral [29]. Both Machine Learning and Natural Language Processing techniques are incorporated in Sentiment Analysis.

Sentiment analysis is currently used in detecting the opinions of social media users. Opinions are studied in several areas such as marketing and politics

2.5. Common Sense Reasoning/Sentic Computing

Common sense is the knowledge (usually acquired in early stages of life) concerning social, political, economic and environmental aspects of the society we live in. Common sense usually varies among different cultures and is built from layers of learning experiments we acquire throughout life [30].

Computers do not have common sense reasoning by nature, but there is a research field, known by Sentic Computing [31], that aims towards transforming computers into machines that could feel. This field of research is a multidisciplinary approach to opinion mining and sentiment analysis, which uses common sense reasoning and web semantics in order to inspect the emotions, not just the opinions from certain text. The term 'Sentic' derives from the Latin 'Sentire', the root of words like sentiment and sensations.

2.6. Performance Measures

Some evaluation metrics were adapted in Information Retrieval (IR) and then extended to other fields of computer science such as ML. These evaluation metrics are used as measures for the performance of IR and ML systems. The most widely used metrics are Recall, Precision, Area under the ROC and F-Measure.

- Recall is the proportion of returned documents (or values) which are relevant (or correct) $Rl \cap Rt$ [32] out of all relevant documents returned and not returned [33]. The metric is also known as Sensitivity of a system.

$$R=(Rl\cap Rt)/Rl \quad (1)$$

- Precision is the proportion of returned documents (or values) which are relevant (or correct) Rl∩Rt [32]. The metric is also known as Accuracy of a system.

$$P = (Rl \cap Rt)/Rt \quad (2)$$

- F-Measure, proposed by van Rijsbergen in 1979, is a weighted harmonic mean of precision and recall. It is a combination between Recall and Precision metrics, which was introduced to overcome the negative correlation between Precision and Recall [34].

$$F\beta=(1+\beta^2)PR/(\beta^2P+R) \quad (3)$$

- F1 is a special case of F- measure with β =1. β Is a parameter to control the balance between Recall and Precision where $0\leq \beta \leq \infty$. When β is set to 0, it implies giving no importance to recall, when β tends to ∞ then no importance is given to precision, and when $\beta = 1$ then Recall and Precision will be given equal importance [35].

$$F1= 2PR/(P+R) \quad (4)$$

P: Precision
R: Recall
Rt: Returned documents
Rl: Relevant documents

- ROC (Receiver Operating Characteristics) graph [36] –or area- is an analysis technique which had been used originally in medical diagnosis, to be later adopted in Machine Learning evaluation. ROC depicts the tradeoff between True Positive and False Negative rates. ROC areas are categorized roughly according to the values:

0.9-1 = excellent (A)
0.8-0.9 = good (B)
0.7-0.8 = fair (C)
0.6-0.7 = poor (D)
0.5-0.6 = fail (F)

3. Previous Work

As stated previously, the research efforts in cyberbullying covered several areas, including the detection of online bullying when it occurs; reporting it to law enforcement agencies, Internet service providers and others for the purpose of prevention and awareness;

and identifying predators with their victims. No effort was directed towards detecting cyberbullying in Arabic language. This section focuses on the previous work done in the areas of cyberbullying detection, ML, feature extraction and cross language transliteration and translation since those are the main techniques used in the implementation of cyberbullying detection systems.

3.1 Cyberbullying Detection

Most of the research done in detecting cyberbullying constituted of either a filtration software or ML techniques. A filtration software has to be employed by social networking platforms, in order to automatically delete or shade profane words [37] [38] [39]; but the filtration method is first limited by its inability for detecting subtle language harassment and second it has to be manually installed [40].

Most work other than filtration methods employs ML techniques, where old corpora of comments or conversations is crawled, whether from Facebook, Twitter, Formspring (a platform similar to Facebook, popular between teens) or even real conversations of sexual attackers [41]. These corpora are used to feed ML algorithms responsible for detecting cyberbullying attacks by building a classification rule from the training set. The obtained classification rule classifies the testing set comments. Such work was done in [22] where the authors crawled data from Formspring and used the Amazons Mechanical Turk [42] for labelling comments. Then they used the learning methods from the Waikato Environment for Knowledge Analysis (WEKA) toolkit [43] to teach and test the model for classifying comments.

The problem of detecting subtle language cyberbullying attacks was tackled by Dinakar et al [6]. They depended on commonsense reasoning in the detection of cyberbullying content. As an example of the commonsense they used: they considered comments of wearing makeup when subjected on Males might indicate the presence of harassment. They built their datasets from both YouTube and Formspring for both training and testing. They used NB, JRip, J48 and SVM for text classifications. For feature sets they used general features, such as a list of unigrams or profane words, tf-idf weighting scheme, Ortony Lexicon for negative affect, Part-of-speech tags for commonly occurring bigrams, and Label Specific Features including frequently used forms of verbal abuse.

The weighting scheme tf-idf is the product of term frequency and the inverse document frequency in the dataset. It involves multiplying term frequency (tf), that represents the number of times a term occurs in a document, by inverse document frequency (idf), which varies inversely with the number of documents to which a word is assigned [44].

Nahar, Li and Pang [40] employed the tf-idf weighting scheme for building features. In addition they built a network composing of bullies and their victims. The network was used to rank the most active predator and its target. In [45] Dinakar et al. stated that detecting profane language cyberbullying is easier than detecting sarcasm and subtle language attacks. Chayan and Shylaja of [46] enhanced the performance of the cyberbullying detection model by 4%, through looking for comments directed towards peers by using Supervised ML and Logical Regression models. However they didn't detect sarcasm comments. Dadvar et al. [47] state that incorporating user context such as the user's history as a feature for training the cyberbullying detection model increases accuracy of classification; however they didn't include sarcasm detection in their system.

SVM was also used by Yin et al. [48] for classifying posts as containing harassment or not. They used documents from CAW 2 dataset, which included posts from Kongregate, Slashdot and Myspace. For feature selection they incorporated several features;

- *Local features,* they used tf-idf.
- S*entiment features,* such as 1- grams, 2- grams and 3- grams, and they also captured second pronouns.
- *Contextual features* in which they studied both the similarity of a post to other neighboring posts and the cluster of posts neighboring around a certain harassment post.

Capturing sentiment features only didn't perform well so they compared the performance of their system by mixing features. Tf-idf performed better than n-grams and foul words, however, combining tf-idf with contextual and sentiment features achieved an additional enhancement in results in Precision, Recall and F1-measure. A similar work was done by Dadvar et al. [47], who built their feature space from Content-based, User based and Context based features. They also proved that incorporating contextual features such as gender information from the user's profile enhances cyberbullying detection.

Other research efforts were focused around social network profiles, such as [17]. They presented a methodology to detect and associate fake profiles on Twitter social networks to real users. This system had been capable of linking the owners of a fake account on Twitter to a real account for one or more students in a school class; this was a case of a real cyberbullying incident. The system was devised by collecting features from tweets then analyzing the features using various supervised ML techniques included in WEKA. Afterwards the performance among these techniques was compared on True Positive Ratio (TPR), False Positive Ratio (FPR) and Area Under ROC Curve (AUC). Bayzick, Kontostathis and Edwards [49] proposed the BULLYTRACER software which detects cyberbullying in chat rooms 58.63% of the time.

Chen et al. [50] proposed a new model for detection which they named as the Lexical Syntactic Feature-based (LSF) model; it achieved a precision of 98.24% and recall of 94.34%. Their model calculated both a post and a user's offensiveness depending on the ratio of offense appearing in a user's posts. This model detects "strong profanity" in online posts by using lexical analysis methods such as Bag of Words; and subtle language harassment to which the authors referred as "weak profanity". Then the model uses semantic analysis and NLP techniques to analyze the context of sentences by studying the grammatical relations among words. This research was an extension to the work presented in [51] for cyberbullying posts filtration.

Most of the research in cyberbullying did not give importance for the distinction between cyberbullying and cyberagression, but

for Hosseinmardi et al. [52]. They proposed a definition for cyberbullying which is the ***repetition*** of harmful actions using electronic devices over a certain period of time. They stated that most of the work previously done in detecting cyberbullying was actually focusing on detecting cyberagression. They define cyberagression as a single instance of harmful action that if repeated over time would be considered as cyberbullying. They also demonstrated that a Linear SVM classifier can significantly improve the accuracy of identifying cyberbullying to 87%. In addition they incorporated using features other than text such as images for better detection of cyberbullying.

Another study on cyber aggression was done by Nakano et al [53]. They analyzed anonymous and non-anonymous questions and answers from ask.fm. Their study shows that anonymous questions tend to be more aggressive than non-anonymous ones. They also showed that replies to anonymous questions tend to be less aggressive than replies to questions from known profiles.

Potha and Maragoudakis [54] stressed on a window of time in order to study the textual patterns of previous conversations in order of predicting the upcoming actions of a predator. They incorporated time series modelling in their research in addition to SVM for features selection. SVD (Singular Value Decomposition) [55] was used for feature reduction and DTW (Dynamic Time Warping) [56] for matching time series collections.

Fuzzy Logic and Genetic algorithms were also used in cyberbullying detection [57], where a new system was proposed using those two methods. This system's performance was compared across precision, recall and F1 measures. The system achieved better in Accuracy, F1-measure and Recall than previous fuzzy classification methods with 0.87, 0.91 and 0.98 respectively.

3.2 Arabic Language

Work related to Arabic language is scarce due to the complex morphological nature of Arabic. Arabic language is used by around 300 million Arabs around the world, mainly Muslims. It is a script language which is read and written from right to left and it constitutes of an alphabet of 28 letters. Vowels in Arabic are represented by special punctuation marks called Diacritics [58]. There are three variations for Arabic languages going on together. The Classical Arabic which is the language of the Islamic manuscripts -such as the Quran and prayers- and Arab people until Mid-20th century. The Modern Standard Arabic (MSA) which is the formal language used nowadays in schools and news, and it is known by all Arabs. Finally comes the dialects, which are accents for the Arabic language, usually used informally between people. There are around 10 dialects, one for every country - or group of countries [59], [60]. Arabic dialects imply a difference in meanings of words between different countries. We might find some words that are considered profane in one country, while good or ordinary in others, for example the word "Yetqalash" in Yemen is a compliment while in Morocco it is an offensive word [61].

Arabic language is a challenging and complex language due to its nature, where Arabic words do not include capitalization [60]. The morphologic nature of Arabic inflicts a lot of ambiguity, and the Arabic corpus is very scarce. Arabic language is ranked the 7th around the world, and its use over internet is growing vastly [59], thus arose the research interest in Arabic language fields.

An extensive search was performed on available articles and publications and no previous work for cyberbullying detection in Arabic language texts and comments was found. But some papers in the fields of ML and NLP were found. The previous work done in Arabic deals with text preprocessing or text classification.

Ghaleb Ali and Omar [62], proposed a key phrase extraction method that combines several key phrase extraction methods with ML algorithms. The output from the key phrase extraction methods is used as features to the ML algorithms. The ML algorithm in turn classifies the feature as either a key phrase or not. They compared their results by using three ML algorithms: Linear Logistic Regression [63], SVM and Linear Discriminant Analysis [64]. They have proved that SVM gives the best results in key phrases extraction among the three algorithms.

Some work had been done in Arabic named entity extraction, such as Named Entity Recognition for Arabic (NERA) [65] to identify proper names in Arabic documents. NERA used a white-list of named entities and corpora compiled from various sources; its performance was measured across recall, precision and F1. The results were satisfactory; 86.3% 89.2% 87.7% respectively for person named entities.

Filtering for spam emails written in Arabic and English was done by El-Halees [66] on pure English, pure Arabic and mixed collections of emails. Several ML techniques were used, including SVM, NB, k-Nearest Neighbor (k-NN) [67] and Neural Networks. The performance of the system was measured across all three variations and SVM was proved to be best in pure English environment. The system performed less on pure Arabic emails, due to the inflictive nature of Arabic Language. The authors also proved that stemming Arabic words enhances the performance of the classifiers, where NB performed best with 96.78% Recall and 92.42% F1-measure.

Other than emails there are also attempts for detecting spam in social networks, such as Twitter. Such work was done by El-Mawass and Alaboodi in [68]. They elaborated a system to detect spam in Arabic tweets. Their system achieved significant accuracy, precision and recall measures.

Sentiment analysis is one of the text classification categories. Sentiment analysis classifies a certain text as positive, negative or neutral [69]. Sentiment analysis was done by Hamouda [70] on Facebook comments written in Arabic. They built a corpora from 6000 comments sampled from Facebook, preprocessed this corpora, and then applied classifications to determine the sentiment behind the comment. Three classifiers were used: SVM, NB and Decision trees. The best performance was achieved by SVM with 73.4% accuracy. Another attempt for sentiment analysis was done in [71] for Arabic Tweets, their special contribution was in handling Arabizi and dialects. They incorporated NB, SVM and k-NN for classification and the best accuracy was approached by NB.

In [72], Duwairi detected sentiments from dialectical Arabic texts. Two methods were applied for detection. First, by translating dialectical words into MSA, then detecting according to MSA

lexicon. Second method was by detecting dialectical lexicon. NB and SVM classifiers were used to detect both negative and positive polarities. The results obtained showed improvement in Precision, Recall and F-measure upon translating into MSA.

Sentiment analysis was applied on Arabizi also by Duwairi et al [69]. In their system they first converted Arabizi text into Arabic by using their own rule based method. They labeled their data using their crowdsourcing tool [71] and then applied SVM and NB for classification. A comparison between SVM and NB showed that SVM outperformed NB. However better results were achieved when they first eliminated neutral entries from the dataset.

Arabic tweets in Saudi Arabia were analyzed by Alhumoud, Albuhairi and Altuwaijri [73]. They analyzed the tweets using a hybrid approach. Their hybrid approach composed of building a classifier and training it using a one-word dictionary. They compared the results obtained from both their hybrid approach and the supervised learning approach. Two classifiers were used from WEKA, NB and SVM. Their results showed the outperformance of the hybrid approach.

A significant research effort was done on stemming for Arabic language. Stemming is a text preprocessing technique. In stemming words are truncated to obtain their roots [74]. Several stemmers for Arabic are available, including rule-based stemmers such as Khoja's [75] and light stemmers. Light stemmers blindly remove letters from words –affixes and suffices – without prior knowledge of roots [76]. Stemmers are either monolingual or multilingual. Gadri and Moussaoui [77] elaborated a multilingual stemmer. Their stemmer is Language independent and it used the n-gram technique. This stemmer segments words into bigrams, then statistical measures are used to reach the best root. This stemmer was tested against English, French and Arabic. The best success rate (94%) was achieved in small Arabic Datasets. In large datasets, the best results were for English (86, 50%) and the worst for Arabic (67, 66%).

4. Proposed System

As stated previously, one of the purposes of this journal is presenting a solution for the problem of cyberbullying in both English and Arabic languages. As seen in the previous section, there is some work done for detection in English, but none in Arabic Language. Proving the hypothesis that Arabic cyberbullying can be detected was a challenge. Thus in the first stage of the system the focus was on detecting cyberbullying in Arabic language.

Since the proposed system employs ML, a dataset had to be prepared to be used for training and testing the system. Two toolkits were tested for ML, Dataiku DSS and WEKA. The decision was to use WEKA toolkit because it supports Arabic language.

4.1 Data Preparation

In order to train and test the system, a huge amount of data had to be obtained. Thus, the choice was to scrap data from both Facebook and Twitter. This choice was inflicted by the fact that those two social media portals are the most widely used by the Arab nation, especially Arab youth.

For the data acquirement phase two custom tools were built. Those tools were implemented to scrap data from social networks, one for Facebook and the latter for Twitter. Twitter Scrapper was written using PHP while Facebook scrapper incorporated python. Both scrappers connected to a mongo dB server, where all downloaded tweets and messages were stored.

In the process of getting the data from Twitter, the Twitter Scrapper searches for Tweets according to a given location and perimeter. A huge database was accumulated using this tool. In order to get the tweets from the Arab countries we focused on the center of the Middle East Region as a location, and selected an area of 10,000 kilometers radius in each run. Thus the tweets were collected from Lebanon, Syria, Gulf Area and Egypt mainly. The size of the tweets database summed up to 4.93 GB.

Collecting Facebook data was harder, due to the restrictions imposed by Facebook security and privacy measures. In Facebook the data source had to be specified beforehand. It is specified by page IDs. The aim was for pages of public figures and news agency, since those pages include more discussions and interactions between Facebook users. The tool had to be run many times to get data. Each run required that the Facebook Page ID is specified to collect the data from it. Many requests were blocked due to privacy measures on the targeted pages. The Facebook database reached 0.98GB of size, thus the decision was to keep Facebook data for validating the system in a later stage and use the tweets dB for training and testing the system at the current stage.

4.2 Data Labelling and Preprocessing

Before the system training phase, the obtained data had to be cleaned and preprocessed. WEKA was used for this purpose as mentioned above. For the system in hand, version 3.9.1 of WEKA was used since it contains specific packages that will be mentioned later.

The only features included in the first stage were text (the content of the tweet) and language. Several languages were observed in the database, mostly Arabic, English and Turkish. All the tweets of languages other than English and Arabic were discarded. The remaining tweets were separated into two datasets, one having only tweets in Arabic language and the latter English. The dataset with only Arabic language was denoted by ArabicTweets. This Dataset contained 35273 unique tweets after removing all duplicates. While the English Dataset contained 91431 tweets.

Due to the sensitivity of Arabic language and to ensure the correct labelling of sarcastic cyberbullying content, Arabic tweets were labelled manually by adding an extra attribute to the dataset, which is the "bullying". This attribute was assigned to be the class attribute, where labelling will take place and it will be used by WEKA to evaluate the performance of the system. The values of bullying instances are either "yes" or "no". "Yes" is when the tweet contains cyberbullying and "no" otherwise. The server used to preprocess the datasets was a 3.5 GHz CPU and 32GB RAM's windows 10 virtual machine.

4.3 System Training and Testing

Training and classification procedures were done several times for the purpose of reaching the best results. Two models were chosen in the first stage, Naïve Bayes and SVM. Those two models were chosen based on analysis of the previous work done in the field of ML. Researchers reached a conclusion that those are the best two algorithms for text classifications – as mention in section 2 [54, 57, 59, and 61]. In the first model the system was trained using Naïve Bayes and promising results were achieved. The precision was 90.8514 % for the overall system. However this precision is somehow tricky. The dataset contained only 2196 bullying content out of the total 35273. Thus even if in the worst case the system classifies all the tweets as not bullying then it will be still achieving a high precision. Hence the analysis of the performance was shifted towards the actual cyberbullying content which was detected (or in other words classified as "yes"). The system in the first run was capable of detecting 801 out of 2196 actual bullying instances (or tweets) which is a good result and it proves the aim of the paper: Cyberbullying in Arabic is detectable!

In training the second model; the AffectiveTweets package was used [78] and specifically the TweetToSentiStrengthFeatureVector filter. This filter depends on the SentiStrength mechanism developed by Thelwall et al [79]. SentiStrength originally works with English, but it is capable of supporting other languages. SentiStrength works on weights of tweets: positive weights (2 to 5) are given to words denoting positive feelings and negative weights for negative feelings (-2 to -5). 1 and -1 are considered neutral. Then a list of Arabic profane words was collected and weighted manually. The weight given for profane words was (-5) which is the extreme negativity in SentiStrength. Afterwards the English lexicon files used by SentiStrength were replaced with custom built Arabic files including the weighted profane words.

In the process of customizing the SentiStrength files, the phrases added contained the profanity from several Arabic countries. This gave a chance for handling dialectical difference between Arabic countries implicitly. With more effort in including more dialectical varieties of phrases all the dialects in Arabic countries would be covered.

The dataset prepared had to be preprocessed before building the training model. Preprocessing was performed using WEKA. The set of techniques used for preprocessing included applying the TweetToSentiStrengthFeatureVector filter, converting strings to word vectors and normalizing. After that the system was trained using SVM. Training and testing the model is a time-consuming process (it took around 8 hours). As a preliminary stage, the system achieved results somehow close to the results achieved by Naïve Bayes in the first model. The model was capable of detecting 710 bullying instances. Those results are promising and they show that SentiStrength can be customized to be used in cyberbullying detection. The results achieved by the system are summarized in the next section.

5. Results Obtained

Table 1 shows the division of the classified instances. How many instances were classified correctly and how many wrong in each model and class. As seen, in Naïve Bayes model 31245 non-bullying instances were classified correctly. There are also 1832 non-bullying instances classified as containing bullying. As for the bullying instances, it is seen that 801 were classified correctly while 1395 were classified as not containing bullying.

Table 1: Classification Results

	Correctly Classified		Miss Classified	
Class	***no***	***yes***	***no***	***yes***
Naïve Bayes	31245	801	1832	1395
SVM	32479	710	1923	161

Table 2: Summary OF Rates

	Class	Precision %	Recall %	TP %	FP %	F-Measure %
Naïve Bayes	***no***	94.5	95.7	95.7	69.6	95.1
	yes	36.5	30.4	30.4	4.3	33.2
	Overall	90.1	90.9	90.9	64.7	90.5
SVM	***no***	94.4	99.5	99.5	73	96.9
	yes	81.5	27	27	0.5	40.5
	Overall	93.4	94.1	94.1	67.6	92.7

Table 2 displays the results obtained by WEKA after training and testing the model. Precision, Recall, True Positive Rate, False Positive Rate and F-Measure are displayed for both models, Naïve Bayes and SVM. In each model the measures mentioned are displayed for the two classes of the "bullying" attribute which are "yes" and "no". In the third row of every model there is the weighted measure from both classes.

There is a difference between the "yes" precisions of the two models noted in Table 2. SVM achieved a much higher precision for the "yes" class -0.815 for SVM and 0.365 for NB. This difference is due to the lower number of wrong classified instances by SVM as seen in Table 2.

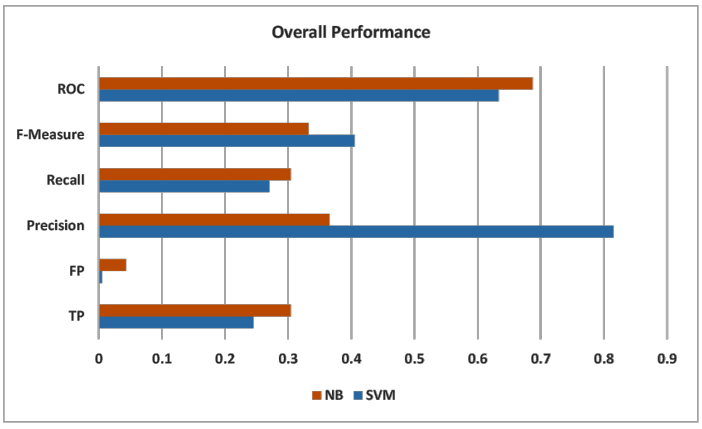

Figure 1: Difference in performance between NB and SVM

Figure 1 displays graphically the difference in performance between NB and SVM for the weighted class, or the overall per-

formance. As seen the precision achieved by SVM was much better than NB, while the other measures were somehow close, such as ROC, TP, FP, etc.

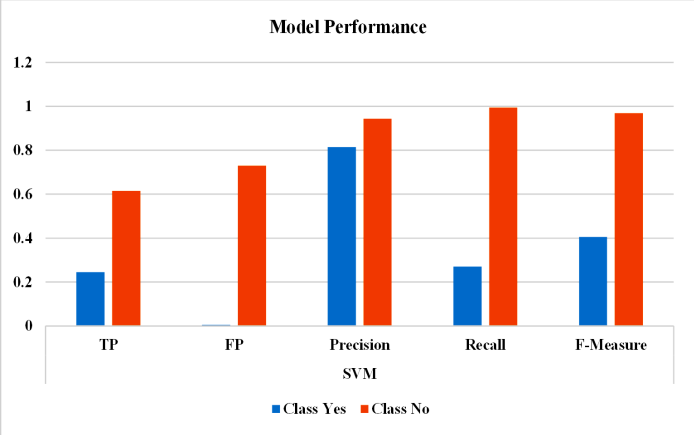

Figure 2: Comparison between the measures achieved by SVM model

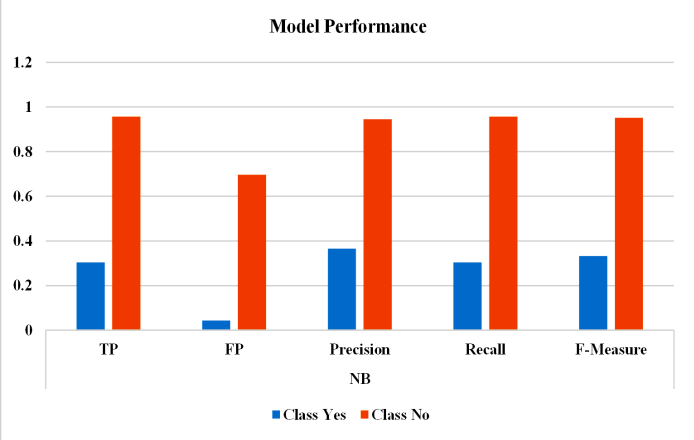

Figure 3: Results of NB Model

Figure 2 compares graphically between the measures achieved in both classes by SVM model, while Figure 3 displays the results of NB. As shown previously Table 1 and Table 2, the low number of False Positives, high precision and recall are significant.

The results obtained by this system are not perfect if compared with previous work done in English cyberbullying detection systems. But the aim of the paper was to prove that Arabic cyberbullying is detectable. Thus, the results displayed in the Tables and Figures above cannot be compared with previous work done for Arabic Language due to the negligibility of such work. Table 2 shows a recall of 30.4 for the "yes" class instances in the NB model, which means at least one third of the bullying in Arabic is detectable by the system.

The overall measures for both models give reasonable results. Although the focus was on the analysis of the "yes" class since it contains the bullying instances which are less than the "no" count. But the high precisions approached by the system denote that the model is not classifying a high number of non-bullying instances as bullying ones. Which is also an important point to consider.

6. Conclusion

The work done in this paper proves that detecting Arabic cyberbullying is possible. However, some effort is yet needed to enhance the performance of the system presented. Therefore, some future plans still need to be completed in later stages.

The first step is to enhance the performance measures achieved by the system through using hybrid training models, such as combinations of Distance Functions, NB and SVM. Treatment of Arabizi content is also in the future plans. This is a crucial point, since a lot of Arabic youth use the chat language on social networks for communication among each other. Thus the proposed system will be upgraded to handle cyberbullying when written in Arabizi.

Further plans incur training the system using deep learning methods instead of machine learning and then comparing the differences of the outcomes from the two schemes.

References

[1] B. Haidar, M. Chamoun and F. Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," in UKSIM 2016, Pisa, Italy, 2016.

[2] K. Poels, A. DeSmet, K. Van Cleemput, S. Bastiaensens, H. Vandebosch and I. De Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders," Cyberbullying on social network sites, vol. 31, p. 259–271, 2014.

[3] S. S. Kazarian and J. Ammar, "School Bullying in the Arab World: A Review," The Arab Journal of Psychiatry , vol. 24, no. 1, pp. 37 - 45, 2013.

[4] M. Y. Ba-Isa, "Beaten up in cyberspace," 25 8 2010. [Online]. Available: http://www.arabnews.com/node/353552. [Accessed 16 9 2016].

[5] ICDL, "Cyber Safety Report: Research into the online behaviour of Arab youth and the risks they face," ICDL Arabia, 2015.

[6] K. DINAKAR, B. JONES, C. HAVASI, H. LIEBERMAN and R. PICARD, "Common Sense Reasoning for Detection, Prevention,and Mitigation of Cyberbullying," in ACM Transactions on Interactive Intelligent Systems, NY, September 2012.

[7] O. f. V. o. C. National Crime Prevention Council, "Cyberbullying Tip Sheets," National Crime Prevention Council, 2016. [Online]. Available: http://www.ncpc.org/topics/cyberbullying/cyberbullying-tip-sheets/. [Accessed 10 June 2016].

[8] N. Willard, "Educator's Guide to Cyberbullying and Cyberthreats," Center for Safe and Responsible Internet Use, 2007.

[9] N. Samaneh, A. Masrah, M. Azmi, M. S. Nurfadhilna, A. Mustapha and S. Shojaee, "13th International Conrence on Intelligent Systems Design and Applications (ISDA)," in A Review of Cyberbullying Detection . An Overview, 2013.

[10] D. Mann, "Emotional Troubles for 'Cyberbullies' and Victims," WebMD Health News, 6 July 2010. [Online]. Available: http://www.webmd.com/parenting/news/20100706/emotional-troubles-for-cyberbullies-and-victims. [Accessed 24 August 2015].

[11] T. M. Mitchell, "The Discipline of Machine Learning," CMU-ML-06-108, Pittsburgh, July 2006.

[12] P. Kulkarni, Reinforcement And Systemic Machine Learning For Decision Making, New Jersey: IEEE, WILEY, 2012.

[13] P. FLACH, MACHINE LEARNING The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.

[14] D. Vilariño, C. Esteban, D. Pinto, I. Olmos and S. León, "Information Retrieval and Classification based Approaches for the Sexual Predator Identification," Faculty of Computer Science, Mexico.

[15] H. José María Gómez and A. A. Caurcel Diaz, "Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification," 2012.

[16] A. S. a. S. Vishwanathan, Introduction to Machine Learning, Cambridge: Cambridge University Press, 2008.

[17] I. Santos, P. G. Bringas, P. Gal´an-Garc´ıa and J. Gaviria de la Puerta, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying," DeustoTech Computing, University of Deusto, 2013.

[18] I.-S. Kang, . C.-K. Kim, . S.-J. Kang and S.-H. Na, IR-based k-Nearest Neighbor Approach for Identifying Abnormal Chat Users, 2012.

[19] C. M. a. G. Hirst, Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features, 2012.

[20] D. E. L. a. A. B. Javier Parapar, "A learning-based approach for the identification of sexual predators in chat logs," 2012.

[21] Ron Kohavi and R. Quinlan, "Decision Tree Discovery," 1999.

[22] K. Reynolds, "Using Machine Learning to Detect Cyberbullying," 2012.

[23] S. Ahmad, "Tutorial on Natural Language Processing," Artificial Intelligence (810:161) Fall 2007.

[24] V. Gupta, "A Survey of Natural Language Processing Techniques," vol. 5, 01 Jan 2014.

[25] B. MANARIS, "Natural Language Processing: A Human–Computer Interaction Perspective," vol. 47, no. pp. 1-66, 1998..

[26] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research," IEEE ComputatIonal IntEllIgEnCE magazIne, May 2014.

[27] C. Surabhi.M, "Natural Language Processing Future," in International Conference on Optical Imaging Sensor and Security, Coimbatore, Tamil Nadu, India, July 2-3, 2013.

[28] G. G. Chowdhury, "Natural Language Processing," Annual Review of Information Science and Technology, vol. 37, no. 0066-4200, pp. 51-89, 2003.

[29] S. Dinakar, P. Andhale and M. Rege, "Sentiment Analysis of Social Network Content," in IEEE 16th International Conference on Information Reuse and Integration, 2015.

[30] E. Cambria, Application of Common Sense Computing for the Development of a Novel Knowledge-Based Opinion Mining Engine, University of Stirling, Scotland, UK, 2011.

[31] M. Grassi, E. Cambria, A. Hussain and F. Piazza, "Sentic Web: A New Paradigm for Managing Social Media Affective Information," Cogn Comput (2011) 3:480–489.

[32] W. E. Webber, Measurement in Information Retrieval Evaluation (Doctor of Philosophy), The University of Melbourne, September 2010.

[33] C. J. v. RIJSBERGEN, INFORMATION RETRIEVAL, University of Glasgow.

[34] N. Chinchor, "MUC-4 EVALUATION METRICS," in Fourth Message Understanding Conference, 1992.

[35] Y. Sasaki, "The truth of the F-measure," University of Manchester, 26th October, 2007.

[36] T. Fawcett, "An introduction to ROC analysis," in Pattern Recognition Letters 27 (2006) 861–874.

[37] WatchGuard, "Stop Cyber-Bullying in its Tracks - Protect Schools and the Workplace," WatchGuard Technologies, 2011.

[38] "https://blog.barracuda.com/2015/02/16/3-ways-the-barracuda-web-filter-can-protect-your-classroom-from-cyberbullying/".

[39] "Internet Monitoring and Web Filtering Solutions," PEARL SOFTWARE, 2015. [Online]. Available: http://www.pearlsoftware.com/solutions/cyber-bullying-in-schools.html. [Accessed 2 June 2016].

[40] V. Nahar, X. Li and C. Pang, "An Effective Approach for Cyberbullying Detection," in Communications in Information Science and Management Engineering, May 2013.

[41] "Perverted Justice," Perverted Justice Foundation, [Online]. Available: http://www.perverted-justice.com/.

[42] "Amazon Mechanical Turk," 15 August 2014. [Online]. Available: http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanical-TurkGettingStartedGuide/ SvcIntro.html. [Accessed 2 June 2016].

[43] S. Garner, "Weka: The waikato environment for knowledge analysis," New Zealand, 1995.

[44] "tf-idf: A single Page Tutorial," [Online]. Available: http://www.tfidf.com. [Accessed 13 May 2016].

[45] K. Dinakar , R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," Cambridge, 2011.

[46] V. S. Chavan and Shylaja S S , "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," in International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[47] M. Dadvar, D. Trieschnigg, R. Ordelman and F. De Jong, "Improving cyberbullying detection with user context," 2013.

[48] D. Yin, Z. Xue, L. Hong, B. D. Davidson, A. Kontostathis and L. Edwards, "Detection of Harassment on Web 2.0," Madrid, Spain, April 21, 2009.

[49] J. Bayzick, A. Kontostathis and L. Edwards, "Detecting the Presence of Cyberbullying Using Computer Software," Koblenz, Germany, June 14-17, 2011.

[50] Y. Chen, S. Zhu, Y. Zhou and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," 2012.

[51] Z. Xu and S. Zhu, "Filtering Offensive Language in Online Communities using Grammatical Relations," Redmond, Washington, US, July 13-14, 2010.

[52] H. Hosseinmardi, S. Arredondo Mattson, R. IbnRafiq, R. Han, Q. Lv and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2015.

[53] T. Nakano, S. Tatsuya, Y. Okaie and M. J. Moore, "Analysis of Cyber Aggression and Cyber-bullying in Social Networking," in IEEE Tenth International Conference on Semantic Computing, 2016.

[54] N. Potha and M. Maragoudakis, "Cyberbullying Detection using Time Series Modeling," 2014.

[55] K. Baker, "Singular Value Decomposition Tutorial," 2013.

[56] M. Muller, "Dynamic Time Warping," in Information Retrieval for Music and Motion, Springer, 2007, pp. 69 - 84.

[57] B. Nandhinia and J. Sheebab , "Online Social Network Bullying Detection Using Intelligence Techniques," 2015.

[58] M. A. Attia, Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation, Doctor of Philosophy in the Faculty of Humanities, 2008.

[59] K. Darwish and W. Magdy, "Arabic Information Retrieval," vol. 7, no. 4, 2013.

[60] A. FARGHALY and K. Shaalan, "Arabic Natural Language Processing:Challenges and Solutions," vol. 8, December 2009.

[61] "12 Arabic Swear Words and Their Meanings You Didn't Know," [Online]. Available: http://scoopempire.com/swear-words-meanings-around-middle-east/#.V0fdjPl96M9. [Accessed 2 June 2016].

[62] N. Ghaleb Ali and N. Omar, "Arabic Keyphrases Extraction Using a Hybrid of Statistical and Machine Learning," in International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia, 2014.

[63] T. Haifley, "Linear Logistic Regression: An Introduction," IEEE, 2002.

[64] G. J. McLACHLAN, "Discriminant Analysis and Statistical Pattern Recognition," Wiley InterScience, New Jersey, 2004.

[65] K. Shaalan and H. Raza, "Arabic Named Entity Recognition from Diverse Text Types," Berlin Heidelberg, GoTAL 2008.

[66] A. El-Halees, "Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques," The International Arab Journal of Information Technology, vol. 6, no. 1, 2009.

[67] T. M. COVER and P. E. HART, "Nearest Neighbor Pattern Classification," IEEE TRANSACTIONS ON INFORMATION THEORY, vol. 13, no. 1, 1967.

[68] N. EI-Mawass and S. Alaboodi, "Detecting Arabic Spammers and Content Polluters on Twitter," in Sixth International Conference on Digital Information Processing and Communications (ICDIPC), 2016.

[69] R. M. Duwairi, M. Alfaqeh, M. Wardat and A. Alrabadi, "Sentiment Analysis for Arabizi Text," in 7th International Conference on Information and Communication Systems (ICICS), 2016.

[70] A. E.-D. A. Hamouda and F. E.-z. El-taher, "Sentiment Analyzer for Arabic Comments System," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no. 3, 2013.

[71] R. M. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment Analysis in Arabic Tweets," in 5th International Conference on Information and Communication Systems (ICICS), 2014.

[72] R. M. Duwairi, "Sentiment Analysis for Dialectical Arabic," in 6th International Conference on Information and Communication Systems (ICICS), 2015.

[73] S. Alhumoud, T. Albuhairi and M. Altuwaijri, "Arabic sentiment analysis using WEKA a hybrid learning approach," in 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, Purtogal, 2015.

[74] A. Al-Zyoud and W. A. Al-Rabayah, "Arabic Stemming Techniques: Comparisons and New Vision," in Proceedings of the 8th IEEE GCC Conference and Exhibition, Muscat, Oman, 2015.

[75] S. Khoja and R. Garside, "Stemming arabic text," Computing Department, Lancaster University, Lancaster, UK, 1999.

[76] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in Arabic Computational Morphology, book chapter, , , Springer, 2007.

[77] S. Gadri and A. Moussaoui, "Information Retrieval: A New Multilingual Stemmer Based on a Statistical Approach," in 3rd International Conference on Control, Engineering & Information Technology (CEIT), 2015.

[78] S. M. Mohammad and F. Bravo-Marquez, "Emotion Intensities in Tweets," in Joint Conference on Lexical and Computational Semantics (*Sem), Vancouver, Canada, August 2017.

[79] M. B. K. P. G. C. D. &. K. A. Thelwall, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, p. 2544–2558, 2010.

[80] Hewlett-Packard Development Company. L.P., 2013. [Online]. Available: http://www.autonomy.com/html/power/idol-10.5/index.html. [Accessed 2 June 2016].

[81] "Arabic chat alphabet," 23 May 2016. [Online]. Available: https://en.wikipedia.org/wiki/Arabic_chat_alphabet. [Accessed 2 June 2016].