A survey on author profiling, deception, and irony detection for the Arabic language

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/1837180 since 2022-01-31T01:32:58Z

Published version:
DOI:10.1111/lnc3.12275

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
A Survey on Author Profiling, Deception and Irony Detection for the Arabic Language
Paolo Rosso¹, Francisco Rangel¹, Irazú Hernández Farías², Leticia Cagnina³, Wajdi Zaghouani⁴, Anis Charfi⁴

¹ Universitat Politècnica de València
² INAOE Puebla
³ Universidad Nacional de San Luis
⁴ Carnegie Mellon University Qatar

Abstract
The possibility of knowing people traits on the basis of what they write is a field of growing interest named author profiling. To infer a user's gender, age, native language, language variety or even when the user lies, simply by analyzing her texts, opens a wide range of possibilities from the point of view of security. In this paper, we review the state of the art about some of the main author profiling problems, as well as deception and irony detection, especially focusing on the Arabic language.

1. Introduction
Idiosyncrasy inherent to social media makes them a special environment of communication, due to its freedom of expression, informality and spontaneous generation of topics and trends. But also the possible anonymity of the users. In most cases, personal information is missing; in other cases, users lie. This lack of knowledge about who writes the contents contributes to the emergence of new security issues, such as threatening messages. For example, Magdy et al. (2015) collected a corpus of approximately 57 thousand tweeps, who authored nearly 123 million tweets. The tweets collected were mostly written in Arabic, and they are related to ISIS organization. To study the historical timeline of the users, Magdy et al. (2015) classified manually the tweets into anti-ISIS, and pro-ISIS, in other words ISIS supporter’s vs ISIS opponents as they checked their historical timeline for the period before the creation of ISIS so they get insights into the antecedents of their support preference. Finally, a classifier was built using the collected data to predict eventually who is more likely will oppose or support the group.

To be able to determine the linguistic profile of a person who writes a "suspicious or threatening text" may provide valuable background information. For example, when analyzing a threatening text, we should know: i) the veracity of the threat, by detecting possible deception or irony in the message (since therefore does not represent a threat); ii) the demographics of the author, such as age, and gender; iii) besides her cultural and social context (e.g. native language or/and dialect), with the

¹ Fake terroristic threat: two Irish were refused entry to the USA after tweeting that they were going to “destroy” America http://abcnews.go.com/Blotter/pair-held-twitter-homeland-threat-mix-reports/story?id=15472918
attempt of profiling potential terrorists (Russell and Miller, 1977). Recently, we started the Arabic Author Profiling project for Cyber-Security (ARAP) to address the lack of resources and tools for the author profiling task in Arabic.

In this survey, we review the state of the art of some of the main author profiling areas in general and for the Arabic language in particular. We focused mainly on Arabic language, in order to stress the gap of what has been addressed in English and other languages as compared to the Arabic language. We start our survey with the age and gender identification task, the native language and language variety identification task. Later on, we present the work on the deception detection and the irony and sarcasm detection. Finally, we briefly discuss some of the challenges faced while processing the Arabic language in these tasks.

2. Age and Gender Identification

Author profiling is a research topic that is in vogue in the research community and several are the shared tasks organized on different demographic aspects during the last years. With respect to age and gender identification, a shared task has been organized at PAN\(^3\) at the Conference and Labs of the Evaluation Forum (CLEF)\(^4\) since 2013. The focus has been on age and gender identification, in different languages apart from English:

- In 2013 (Rangel et al., 2013), the aim was dealing with large datasets with high levels of noise, both in English and Spanish.
- In 2014 (Rangel et al., 2014), participants had to approach the task in multiple genres such as social media, blogs, Twitter and hotel reviews, for both English and Spanish.
- In 2015 (Rangel et al., 2015), age and gender identification problem was combined with personality recognition. In this case, the tweets were provided for Spanish, English, Italian and Dutch.
- In 2016 (Rangel et al., 2016b), the focus was on the cross-genre evaluation, that is, training in one genre (Twitter) and evaluating in another one (blogs, social media and reviews). This year data was provided for Spanish, English and Dutch.
- In 2017 (in progress), the goal is to identify the authors’ gender as well as the specific variation of their native language.

Majority of approaches at PAN used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, and so on, together with parts-of-speech tags and content-based features such as bag of words, term frequency-inverse document frequency (TF-IDF), dictionary-based words, topic-based words, entropy-based words, or content-based features obtained with Latent Semantic Analysis (LSA). It should be highlighted this approach that obtained the overall best results for three years (López-Monroy et al., 2013; López-Monroy et al., 2014; Álvarez-Carmona et al., 2015) by using a second-order representation that relates documents with profiles (e.g. men, women, teenagers, etcetera) and subprofiles (e.g. videogamers, students, housewives, etc.). In another work (López-Monroy et al., 2015), the authors test their approach on Schler’s collection (Schler et al., 2006) showing a significant improvement in

\(^{2}\) [http://arap.qatar.cmu.edu/](http://arap.qatar.cmu.edu/)
\(^{3}\) [http://pan.webis.de](http://pan.webis.de)
\(^{4}\) [http://www.clef-initiative.eu](http://www.clef-initiative.eu)
accuracy up to 82.01% and 77.68% respectively for gender and age identification. On the English partition of the PAN-AP-13 dataset (Rangel et al., 2013), the authors in (Weren et al., 2014) show the contribution to the task of information retrieval features, obtaining accuracies of 62.1% and 68.2% respectively for gender and age identification. The authors in (Maharjan et al., 2014) approach the task with 3 million features processed with MapReduce, that allow them to obtain competitive results (higher than 61% for both gender and age identification in both English and Spanish datasets) with great reductions in time consumed. Finally, the EmoGraph graph-based approach (Rangel and Rosso, 2016) captures how users convey verbal emotions in the morphosyntactic structure of the discourse, obtaining competitive results with the best-performing systems at PAN 2013 and demonstrating its robustness against genres and languages on PAN-AP-14 corpus (Rangel and Rosso, 2015).

2.1 Age and Gender Identification in Arabic

The literature for age and gender identification in the Arabic language is scanty. The authors in (Estival et al., 2008) investigate the age and gender identification problem (besides the level of education or personality) in English and Arabic emails. For Arabic, they collect 8,028 emails from 1,030 native speakers of Egyptian Arabic. They built the Text Attribution Tool (TAT) by obtaining 518 features grouped as shown in Table 1, and test different machine learning algorithms such as support vector machines (SVM), k-nearest neighbors (KNN) or decision trees combined with chi-square or information gain. The accuracies reported are of 72.10% and 81.15% respectively for gender and age identification.

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArabicNamedEntities</td>
<td>Language-independent named entities</td>
</tr>
<tr>
<td>ArabicChar</td>
<td>Character level features</td>
</tr>
<tr>
<td>ArabicMorphological</td>
<td>Morphological level features</td>
</tr>
<tr>
<td>ArabicLexical</td>
<td>Lexical level features</td>
</tr>
</tbody>
</table>

Table 1: Feature groups for the TAT system.

The TAT system includes several data repositories and a couple of components to derive the features and to build classifiers. The architecture is modular and it is organized around a chain of processing modules. This architecture allows a flexible experimentation with the different modules. As shown in Figure 1, The process is data-driven as the output of each processing module depends on its input (Estival et al., 2008).
The authors in (Alsmearat et al., 2015) investigate gender identification in 500 articles collected from well-known Arabic newsletters. They collect articles from writers with similar academic profiles and with experience in journalistic writings and who write their articles in Modern Standard Arabic (MSA), from the Jordan and Palestine variations. They combine bag-of-words features with sentiments and emotions and explore different machine learning methods. In Table 2 their best results are shown. Subsequently, the authors (Alsmearat et al., 2014) extend their work to experiment with different machine learning algorithms, data-subsets and feature selection methods, reporting accuracies up to 94%.

<table>
<thead>
<tr>
<th>Approach Used</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words</td>
<td>86.4%</td>
</tr>
<tr>
<td>Sentiments &amp; emotions</td>
<td>61.9%</td>
</tr>
<tr>
<td>Both</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

Table 2: Results for Alsmearat et al. (2015) in Arabic newsletters.

The authors in (AlSukhni and Alequr, 2016) collect 8,034 tweets from Jordanian dialects and label them manually with gender. They add to their bag-of-words approach the name of the authors of the tweets, reporting a great improvement in different evaluation metrics. They also add other features such as the number of words per tweet or the average word length. Several different machine learning algorithms are tested and the best results are shown in Table 3.

<table>
<thead>
<tr>
<th>Approach Used</th>
<th>Results</th>
</tr>
</thead>
</table>

Figure 1: TAT System Diagram (Estival et al., 2008)
3. Native Language, Language Varieties, and Dialects Identification

Besides the language identification of a potentially threatening message, and especially with the rise of social media, there are new challenges to deal with such as the identification of the native language of its author or even the discrimination among varieties of the same language and dialects.

Native language identification consists of identifying the native language (L1) of an author who writes in another language (L2). This task is crucial for security because it allows contextualizing the author of a possible threat. For example, an author can be writing in Arabic albeit his native language may be Farsi or French, because he was born in France.

Several corpora have been built, mainly from academia where English is learned as a second language. For example, the two versions of the International Corpus of Learner English (ICLE & ICLEv2) (Granger et al., 2002), First Certificate in English (FCE) (Yannakoudakis et al., 2011), International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011), Test of English as a Foreign Language (TOEFL) (Blanchard et al., 2013), International Corpus of Cross linguistic Interlanguage (ICCI) (Tono, 2012), National University of Singapore Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013), Corpus of English Essays by Asian University Students (CEEAUS) (Ishikawa, 2009). Similarly, Lang-8 is a collaborative service where students from different languages can write essays to be corrected by native speakers.

Due to the interest in the field, the first shared task on native language identification was organized at the Innovative Use of NLP for Building Educational Applications (BEA-8) workshop at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HT (Tetreault et al., 2013). There were 29 participants who had to discriminate among 11 languages of the TOEFL corpus. The most used features were character, word and POS n-grams, with support vector machine, maximum entropy and ensemble methods. The reported accuracies are approximately 84%.

On the other hand, the task of discriminating among similar languages such as Bosnian, Croatian and Serbian, or language varieties such as Portuguese from Brazil vs. Portugal, or Spanish from Spain vs. Argentina or Peru, steps up the difficulty of native language identification due both to the highest lexical, syntactical and semantic similarity of the texts, and the cultural idiosyncrasies of the writers.

This field has attracted the researcher’s attention during the last years. There are several investigations with different languages such as English

### Table 3: Results for Alsukhni et al. (2016) in Twitter.

<table>
<thead>
<tr>
<th>Method Description</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words</td>
<td>62.49%</td>
</tr>
<tr>
<td>Bag-of-words + author’s names</td>
<td>98.69%</td>
</tr>
<tr>
<td>Bag-of-words + number of words &amp; average word length</td>
<td>99.50%</td>
</tr>
</tbody>
</table>

6 [https://sites.google.com/site/nlisharedtask2013](https://sites.google.com/site/nlisharedtask2013)
(Lui and Cook, 2013), South-Slavic (Ljubesic et al., 2007), Chinese (Huang and Lee, 2008), Persian and Dari (Malmasi et al., 2015), or Malay and Indonesian (Bali, 2006), to mention just a few of them. For example, focusing on Portuguese the authors in (Zampieri and Gebre, 2012) collect 1,000 articles from well-known Brazilian7 and Portugal8 newsletters. They combine character and word n-grams and report accuracies of 99.6% with word unigrams, 91.2% with word bigrams and 99.8% with character 4-grams. With respect to Spanish, the authors in (Maier and Gómez-Rodríguez, 2014) investigate the identification among Argentinian, Chilean, Colombian, Mexican and Spanish on Twitter. They combine four types of features (n-grams and language models) and report accuracies of about 60-70%. The authors in (Rangel et al., 2016a) collect the HispaBlogs9 corpus by gathering posts from five Spanish varieties: Argentinian, Chilean, Mexican, Peruvian and Spanish. The authors ensure that training and test partitions do not share any author of instance between them, avoiding possible over-fitting. A low-dimensionality representation is proposed to reduce the number of features to only six per class, allowing to deal with big data environments such as social media. They report an accuracy of 71.1% in comparison to 72.2% and 70.8% that they obtain with Skip-grams and Sentence Vectors in (Franco-Salvador et al., 2015).

The interest in the field is also reflected in the number of workshops and shared tasks organized:

- Defi Fouille de Textes (DEFT) 2010 shared task (Grouin et al., 2011) focused on language variety identification of French texts.
- LT4CloseLang workshop on Language Technology for Closely Related Languages and Language Variants (Nakov et al., 2014) organized in 2014 at the conference on Empirical Methods on Natural Language Processing (EMNLP)10.
- VarDial Workshop (Zampieri et al., 2014) on applying NLP Tools to Similar Languages, Varieties and Dialects, organized in 2014 at the International Conference on Computational Linguistics (COLING)11, focused on thirteen languages divided into the following groups: Bosnian, Croatian, Serbian; Indonesian, Malay; Czech, Slovak; Brazilian Portuguese, European Portuguese; Peninsular Spanish, Argentinian Spanish; and American English, British English.
- LT4VarDial joint workshop on Language Technology for Closely Related Languages, Varieties and Dialects (Zampieri et al., 2015) organized in 2015 at RANLP12, focused on thirteen languages divided into the following groups: Bulgarian, Macedonian; Bosnian, Croatian, Serbian; Czech, Slovak; Malay, Indonesian; Brazilian, European Portuguese; Argentinian, Peninsular Spanish; and a group with a variety of other languages.

7 http://www.folha.uol.com.br
8 http://www.dn.pt
9 https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs
10 http://alt.qcri.org/LT4CloseLang/index.html
11 http://corporavm.uni-koeln.de/vardial/sharetask.html
12 http://ttg.uni-saarland.de/lt4vardial2015/dsl.html
Vardial workshop on NLP for Similar Languages, Varieties and Dialects (Malmasi et al., 2016) organized in 2016 at COLING\(^\text{13}\), with two subtasks: \(i\) a more realistic DSL (Discriminating Similar Languages) task with new varieties such as Hexagonal vs. Canadian French, and the removal of very easy to discriminate languages such as Czech vs. Slovak and Bulgarian vs. Macedonian; and \(ii\) a new subtask on discriminating Arabic dialects in speech transcripts (Ali et al., 2015) with Modern Standard Arabic and four dialects (Egyptian, Gulf, Levantine and North African), as described more in detail in Section 3.2.

Author Profiling at PAN 2017, where together with gender identification, the aim is to detect the language variety of the authors. Four are the addressed languages with different variations: \(i\) English (Australia, Canada, Great Britain, Ireland, New Zealand, United States); \(ii\) Spanish (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela); \(iii\) Portuguese (Brazil, Portugal); and \(iv\) Arabic (Egypt, Gulf, Levantine, Maghrebi). For each variety, there are 1,000 authors (half per gender) with 100 tweets per author.

### 3.1 Arabic Native Language Identification

Few are the resources available for the Arabic language. It is worth to mention the BUiD Arab Learner Corpus (BALC) (Randall and Groom, 2009), a resource for studying the acquisition of English spelling. BUiD is a set of examination essays written by 16-year-old Arabic students with different proficiency levels in English. The corpus consists of 1,865 texts with 287,227-word tokens and 20,275-word types. The aim of this research project carried out in collaboration by the British University in Dubai, the United Arab Emirates, and the University of Birmingham in the UK, is to study the particular difficulties for Arab learners when spelling English. The authors draw some preliminary findings consistent with previous studies (Haggan, 1991; Sadhwani, 2005): Arab readers and writers have more problems with vowels than with consonants, reflecting the fact that Arabic is a consonantal script hence Arabs may pay more attention to consonants than to vowels (vowel blindness) (Hayes-Harb, 2006; Ryan and Meara, 1992).

Alfaifi et al. (2014) created the Arabic Learner Corpus (ALC), a large Arabic learner corpus (282K words) produced by native and non-native learners of Arabic from pre-university and university levels. Farwaneh and Tamimi (2012) built the Arabic Learners Written Corpus (ALWC). The corpus of 51K words was produced by non-native Arabic speakers in various countries over a period of 15 years. The corpus covers three basic learner's levels (beginner, intermediate and advanced), and three text styles (descriptive, narrative and instructional). Abuhakema et al. (2008) created a corpus of 9K Arabic words written by native English speakers who learned Arabic as a foreign language while studying abroad. Hassan and Daud (2011) built the Malaysian Arabic Learners Corpus, they tried to investigate the usage of Arabic conjunctions among L2 learners. The corpus size is 240K words and it was written by Malaysian university students during their first and second year of Arabic major degree. Moreover, the corpus includes spontaneous essays produced using Microsoft Word.

\(^{13}\) [http://ttg.uni-saarland.de/vardial2016](http://ttg.uni-saarland.de/vardial2016)
Zaghouani et al. (2015), within the scope of the Qatar Arabic Language Bank (QALB) project (Zaghouani et al. 2014), created a corpus of 2 million words of spelling errors for a variety of Arabic texts including (a) user comments on news websites, including dialectal Arabic (b) native speaker essays (c) non-native speaker essays, (d) machine translation output. The native student essays data is categorized by the student learning level (beginner, intermediate, advanced) and by the learner type (L2 vs. L1).

The goal of the automatic native language identification tool is to find the native language of the language learner using his writing. Most of the research in this area has been done on the native language of English learners. Recently, some efforts were made to identify the native language of text written in other languages such as Arabic. Malmasi and Dras (2014) built an SVM model using various features including function words, part-of-speech n-grams, and Context-Free Grammar (CFG) rules. Their system obtained an accuracy of 41% when it was evaluated using the Arabic learner corpus created by Alfaifi et al. (2014). More recently, Ionescu (2015) created a new distance measure for strings with the name, Local Rank Distance (LRD). His method was inspired by the rank distance method as it measures the local displacement of character n-grams among two strings. During the evaluation of the ALC corpus, Ionescu system outperformed Malmasi and Dras by 10 folds with an accuracy of 50.1%. Finally, Mechti et al. (2016), proposed a classification method using some statistical data generated from a corpus. It is considered a hybrid method combining surface analysis in the text with an automatic learning method.

### 3.2 Arabic Dialects Identification

The lack of language resources for dialectical Arabic well known, recently some researchers addressed this problem by creating lexicons, Wordnets, corpora, and treebanks. In (Zaidan and Callison-Burch, 2011) the authors collect the Arabic Online Commentary dataset (AOC), gathering 86.1K articles and 1.4M comments from three newspapers: i) Al-Ghad\(^{14}\) from Jordan; ii) Al-Riyadh\(^{15}\) from Arabia Saudi; and iii) Al-Youm Al-Sabe’\(^{16}\) from Egypt. They use Amazon Mechanical Turk to manually label them with the corresponding dialect. With a smoothed word unigram model (Zaidan and Callison-Burch, 2014), they report accuracies of 87.2%, 83.3% and 87.9% respectively for Levantine, Gulf and Egyptian dialects. Also, in (Cotterell & Callison-Burch, 2014), a multi-genre dialectal corpus for Levantine, Gulf, North African, Iraqi and Egyptian dialects was described.

Graff et al. (2006), presented an Iraqi Arabic lexicon with words from recorded speech marked with morphology information, pronunciation, and part-of-speech. The annotation was done through a dedicated user interface. Boujelbane et al. (2013), built a Tunisian dialectal corpus in order to create a language model for a speech recognition system for a Tunisian Broadcast News company. Cavalli-Sforza et al. (2013) created an Iraqi Arabic WordNet using an English-Iraqi dictionary and the modern standard Arabic version of

\(^{14}\) [http://www.alghad.com](http://www.alghad.com)
\(^{15}\) [http://www.arlyadh.com](http://www.arlyadh.com)
\(^{16}\) [http://www.youm7.com](http://www.youm7.com)
WordNet as well as the English WordNet. Moreover, a Tunisian dialect WordNet was built in (Bouchlaghem & Elkhifi, 2014) starting from a Tunisian corpus.

Duh & Kirchhoff (2006), built a Levantine lexicon using a transductive learning method through partially annotated text in order to perform sentiment analysis of social networks data using a dedicated lexicon for slang sentimental words and idioms was developed as described in (Hedar & Doss, 2013). Al-Sabbagh & Girju (2012b), described their initial work on building a corpus for Egyptian Arabic. The corpus was compiled from various data sources such as Twitter, Blogs and Forums. Also, Almeman & Lee (2013), used the web as a source to create a multi-dialect Arabic corpus for North African, Egyptian, Gulf and Levantine dialects.

Jarrar et al. (2014) presented his Palestinian Arabic corpus with 43K words and a parallel corpus for Algerian Arabic and MSA was proposed in (Harrat et al., 2014) for the purpose of machine translation. In (Elfardy and Diab, 2013) the authors investigate the discrimination between Egyptian and Modern Standard Arabic. They propose two set of features: i) core features such as token-based, perplexity, morphological-based, orthography, and similar; and ii) meta-features such as frequencies of punctuation signs, numbers, special characters, words in Roman script, words with character flooding, number of words, average word length, and so on. They report an accuracy of 85.5%.

The AOC dataset has been used in other investigations. For example, the authors in (Tillmann et al., 2014) discriminate between Egyptian dialect and Modern Standard Arabic by using a combination of character, word and part-of-speech n-grams with features obtained with the AIDA tool. They report an accuracy of 89.1%. In (Darwish et al., 2014), the authors combine the Egyptian part of the LDC2012T09 dataset (Zbib et al., 2012) with the Modern Standard Arabic dataset of the International Workshop on Arabic Language Translation17. They experiment with different combinations of machine learning features: i) word 1/3-grams with character 1/5-grams, reporting an accuracy of 84.7%; ii) morphological features, reporting accuracies between 89.3% and 90.1%; and iii) the use of a dialectal Egyptian lexicon, reporting accuracies of 93.6% by using 1,300 dialectal words, 94.6% by using 94K verbs and 94.4\% by using 8K words with letter substitutions.

The authors in (Sadat et al., 2014) investigate machine learning techniques using Naïve Bayes classifiers and n-gram Markov language models for the automatic discrimination among 6 Arabic dialects: Egyptian, Iraqi, Gulf (including Bahrain, Emirates, Kuwait, Qatar, Oman and Saudi Arabia), Maghreb (including Algeria, Tunisia, Morocco, Libya and Mauritania), Levantine (including Jordan, Lebanon, Palestine and Syria), and Sudan. They use n-gram models and report accuracies close to 98%.

An interesting work is the one done on Algerian Arabic, Berber and Standard Arabic in (Adouane et al. 2017; Adouane et al. 2016a; Adouane et al. 2016b; Adouane et al. 2016c). The authors used hybrid methods combining dictionaries and supervised machine learning methods such as the Hidden

Markov Model (HMM) and N-gram classification tagging to identify the dialects while Saâdane et al. (2017; Saâdane et al. 2015) used mainly a rule-based linguistic approach to detect the Arabic dialects.

The authors in (Shoufan and Al-Ameri, 2015) provided a comprehensive survey on natural language processing methods for Arabic, including a review of the dialect identification task. The increasing interest in Arabic dialects identification is attested by the eighteen teams participating in the Arabic subtask of the third DSL track (Ali et al., 2015). Its difficulty is backed up by the obtained accuracies of about 50%. A summary of the best approaches and their accuracies is shown in Table 4.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPEN TRACK</td>
<td></td>
</tr>
<tr>
<td>SVM, w/c n-grams</td>
<td>51.4%</td>
</tr>
<tr>
<td>Ensemble, w/c n-grams</td>
<td>51.2%</td>
</tr>
<tr>
<td>Multiple string kernels</td>
<td>50.9%</td>
</tr>
<tr>
<td>CLOSE TRACK</td>
<td></td>
</tr>
<tr>
<td>SVM, char 5/6-grams</td>
<td>53.2%</td>
</tr>
<tr>
<td>Ensemble, w/c n-grams</td>
<td>49.1%</td>
</tr>
</tbody>
</table>

Table 4: Results for 2016 DSL Arabic subtask.

Recently, Habash et al. (2017) proposed MADAR, a large-scale project for dialectal Arabic covering 25 Arabic dialects from the main cities in the Arab region. Their dialectal identification tool is currently in progress.

4. Deception Detection

A deceptive opinion can be defined as a fictitious opinion written with the intention to sound authentic in order to mislead the reader. An opinion spam usually is a short text written by an unknown author using a not very well defined style. These characteristics make the problem of automatic detection of opinion spam very challenging (Rosso and Cagnina, 2017).

In the literature, the most research attends the problem of opinion spam detection for reviews written in English language. In (Fitzpatrick et al., 2015) the authors describe different behaviors indicative of deception such as physiological, gestural and verbal, considering the opinion spam detection among others problems. Several works approached the problem of the detection of deceptive opinions considering features based on the content of the reviews in a supervised way. In (Ott et al., 2011) the authors use the 80 dimensions of LIWC2007 (Pennebaker et al., 2007b), unigrams and bigrams as a set of features with an SVM classifier. In (Hernández Fusilier et al., 2015b; Hernández Fusilier et al., 2015a) the authors propose a PU-learning variant using two different representations: word n-grams and character n-grams. The best results are obtained with a Naïve Bayes classifier using character 4 and 5 grams as features (Hernández Fusilier et al., 2015b) and, the conjunction of word unigrams and bigrams in (Hernández Fusilier et al., 2015a). With those results, the authors conclude that PU-learning show to be appropriate for detecting opinion spam. Character n-grams in tokens, the sentiment score and LIWC linguistic features such as pronouns, articles, and

18 nlp.qatar.cmu.edu/madar
verbs (present, past and future tenses) were used in (Cagnina and Rosso, 2015) for the detection of opinion spam. The best results are obtained with a Naïve Bayes classifier and the combination of character 4-grams in tokens and LIWC features for the representation of the opinions.

**4.1 Deception Detection in Arabic**

The detection of spam in Arabic opinion reviews is a relatively new research field then, the bibliography is scarce in this area. In (Wahsheh et al., 2013b) the authors present one of the first systems to detect spam in Arabic opinions. The system named SPAR uses features as spam URLs (a blacklist with Arabic content/link spam web pages (Wahsheh et al., 2013a)), five or more consecutive numbers and, presence of the '@' symbol with letters around (e-mails address) for the classification of opinions like spam or not spam. The system also categorizes the spam opinions in 'high' or 'low' spam depending on the content of the review, using a special metric. At the same time, the non-spam reviews are labeled as 'positive', 'negative' or 'neutral' based on two language polarity dictionaries built by the authors, one with 2,800 words/phrases and other with 75 emoticons. SPAR is tested with a dataset of 3,090 opinions written in the Arabic language collected manually by the authors from Yahoo!-Maktoob News. An SVM classifier in Weka data mining tool is used to obtain the results. After performing a 10 fold cross-validation experiment, the accuracy reported is 97.50% and the error rate is 2.49%. The authors conclude that SPAR provides a reliable and trustworthy performance to distinguish spam from non-spam opinions.

In (Hammad and El-Halees, 2015) the authors propose a novel approach combining methods from data mining and text mining with machine learning techniques to detect spam in opinion reviews written in the Arabic language. Additionally, the approach uses methods to solve the class imbalance problem present in the dataset used. For the representation of the reviews, review content, meta-data about each reviewer and hotel information have been used as features. The authors build a dataset of 2,848 reviews from online Arabic websites such as Tripadvisor.com.es, Booking.com and Agoda.es. The classification is performed with Naïve Bayes (NB), SVM, ID3 and K-NN algorithms with a 10-fold cross-validation experiment. The best results are obtained with NB and over-sample method, that is 99.20% of accuracy, concluding in the effectiveness of this approach for identifying spam in Arabic reviews.

The authors in (Aloshban and Al-Dossari, 2016) present some preliminary ideas about a method for grouping spam detection in social media for the Arabic language. The proposal uses open source tools for the processing of the Arabic texts and consists of 4 phases: crawling (to collect tweets), preprocessing (to clean the texts), spamming activities detection and individual member's behavior scanning (to identify suspected spammers). The spam activities detection is based on the work of (Mukherjee et al., 2012) that aims to detect group members posting tweets on a particular entity for a short time (Group Time Window), check the similarity of a tweet content (Group Content Similarity) and detect if the members of a group post tweet on the entity at first (Group Early Time Frame). The authors
conclude that the research is at its early stage and a lot of work is still needed to finish this proposal.

5. Irony and Sarcasm Detection

A suspicious message may not be a threat when it is humoristic or ironic (Reyes et al., 2012). Irony and sarcasm represent an interesting way to communicate opinions toward a particular target in social media (Hernández Farias and Rosso, 2016). The most common definition of irony refers to the use of words for expressing the opposite meaning from what is literally said (Grice, 1975). When irony becomes offensive with a specific target to attack is considered as a form of sarcasm (Bowes and Katz, 2011). These figurative language devices represent a big challenge for natural language processing related tasks, especially for sentiment analysis (Bosco et al., 2013).

In recent years, several approaches have been proposed to deal with irony and sarcasm detection in social media. Irony (and sarcasm) detection has been addressed as a classification problem, where decision trees and support vector machine are among the classifiers that obtain the best results. The majority of research investigating irony and sarcasm detection has focused on Twitter. Surface features (such as punctuation marks and emoticons) together with textual markers to identify inconsistencies and incongruities in texts have been widely exploited (Reyes et al., 2013; Barbieri and Saggion, 2014; Hernández Farias et al., 2015; Joshi et al., 2015; Karoui et al., 2015). In (Barbieri et al., 2014; Sulis et al., 2016) the authors attempt to classify tweets labeled with #irony and #sarcasm. They use the same dataset achieving 0.62 and 0.69 in F-measure terms, respectively. Aiming to evaluate the performance of sentiment analysis systems in the presence of irony and sarcasm some evaluation campaigns have been organized in English (Ghosh et al., 2015) and in other languages such as Italian (Basile et al., 2014; Barbieri et al., 2016).

5.1 Irony and Sarcasm Detection in Arabic

With respect to works in Arabic, few are the attempts in which irony has been addressed in literature and mass media (Abuhajam, 2004; Alabban, 2014; Alharbi, 2015; Battish, 1983). There are no automatic approaches to detect irony and sarcasm. In (Sigar and Taha, 2012) the authors manually analyze the similarities and differences between ironic expressions in English and Arabic. They use data from books, articles and Internet (some images). A manual annotated Twitter dataset is instead described in (Refae and Rieser, 2014). The authors asked two native speakers of Arabic to annotate polarity. Additionally, the presence of sarcasm has been annotated. Very recently a preliminary system for irony detection in Arabic in social media was presented in (Karoui et al., 2017). Several features have been taken into account: surface, sentiment, contextual, and shifter ones (e.g. false assertion, exaggeration). In In the future, the authors plan to manually check the reliability of the hashtags they consider and include pragmatic features that should help to infer the context needed to understand further irony.
6. Challenges in Processing Arabic text

Processing the Arabic language for any NLP task can be sometimes challenging due to several peculiarities that we present in this section. First of all, Arabic morphology is relatively complex in that it uses prefixes, infixes and suffixes, not only for inflection but also to concatenate words. This various morphological variation can be dealt with by using hand-crafted rules, which enable to strip off possible prefixes and suffixes from the word stem before further processing. Furthermore, the spoken form of Arabic is quite different from the written form of the language as it is one of the few languages in the world with clear diglossia. For any native speaker of Arabic, there exist at least two forms of the language, the spoken form which is typically a specific dialect versus a formal written form, referred to as modern standard Arabic (MSA). Moreover, Arabic is different from English both morphologically and syntactically. Hence, Arabic is a challenging language to the existing NLP technology tailored to the nuances of the English language. From the morphological standpoint, Arabic exhibits rich morphology. Similar to English, Arabic verbs are marked explicitly for tense, voice and person, however, in addition, Arabic marks verbs with mood (subjunctive, indicative and jussive) information. Depending on the genre of the text at hand, not all of those features are explicitly marked in the naturally occurring text. Arabic writing is known for being underspecified for short vowels.

Developing NLP systems in a diglossic situation like Arabic could indeed lead to some complication. For instance, it is very difficult for any single NLP application to process data from all the dialectal varieties of Arabic with their linguistic peculiarities (e.g. the loss of case distinctions) while they have some common properties. In order to successfully process a text with dialectal Arabic, the NLP application should be able to detect beforehand which variety it is aiming to address so the linguistic properties of the particular dialect can be applied. In order to tackle this issue, Habash et al. (2005) took the initiative to address the issue of Arabic dialects and made the assumption that it is much easier to develop natural language processing tools for the dialects by extracting and categorizing the grammatical features of a given dialect, making it to behave like Modern Standard Arabic (MSA) before applying MSA natural language processing tools to process a text. Currently, the MADAMIRA tool for the morphological analysis and disambiguation of Arabic is widely used and can be considered a state of the art tool to process Arabic (Arfath et al. 2014).

Conclusions

In this survey, we have reviewed the state of the art in the Arabic language of age, gender, native language and language variety identification, as well as of deception and irony detection. The main aim is to highlight what still needs to be done for the Arabic language for automatically profiling demographics or detecting deception and irony. The final aim will be to fill in these gaps and develop an author profiling system for cyber-security in Arabic.
References


Boujelbane, R., BenAyed, S., & Belguith, L. H. (2013). Building bilingual lexicon to create dialect Tunisian corpora and adapt language model. ACL 2013


Granger, G., Dagneaux, E., Meunier, E. (2002). The international corpus of learner English. handbook and CD-ROM.


Analysis, volume 9117 of LNCS, pages 337–344. Springer International Publishing.


