MASHAEIR: Bootstrapping a Multi-Dialect Fine-Grained Emotion Thesaurus for Arabic Using Twitter

KhaledElghamry
Alsun, Ain Shams University, Egypt
elghamryk@gmail.com

Abstract—***The user-generated content on social media sites, e.g. Twitter and Facebook, provides a rich source of people's emotions towards products, issues, people and major events. Accordingly, the focus of more research has moved from negative-positive sentiment classification tasks to tasks of recognizing more fine-grained emotions. However, research on and resources for fine-grained emotion identification in Arabic texts are still lacking. To fill in this gap, this paper introduces MASHAEIR (an Arabic word that means 'emotions'), a corpus-based multi-dialect fine-grained emotion thesaurus for Arabic. MASHAEIR was bootstrapped using 'big data' from Arabic Twitter from January 2007 to July 2015. The thesaurus is enriched with (i) different types of single- as well as multi-word terms expressing emotions, (ii) Arabic dialectal variations in the expression of emotions and (iii) scores that reflect the intensity of the emotions conveyed through these units. The paper also presents a simple evaluation of the thesaurus coverage on a sample Twitter corpus. MASHAEIR is intended to present an outline of a large-scale and easy-to-update emotion thesaurus for Arabic that could also be enriched in the future with more information such as gender and age preferences in expressing emotions.***

Key words: ***Arabic Sentiment Analysis, Emotion Thesaurus, Social Media***

1 INTRODUCTION

With the ever-increasing volume of user-generated content on social media sites like Facebook and Twitter, using the simplistic negative-positive sentiment classification would leave undetected more fine-grained emotions expressed by millions of users of such sites worldwide. Accordingly, there is growing interest in advanced sentiment analysis that would capture more deeply the mood states of social media users, as well as their emotions towards different products, organizations, issues, public figures and major current events, locally and globally (see for example, [3], [4], [12], [31], [10], [15], [18], [23], [27], [32], [11], and [20]).

Recently, there have been efforts to build linguistic resources for Arabic sentiment analysis. For example, [1] reported efforts to build SANA, a large-scale, multi-genre, multi-dialect multi-lingual lexicon for the subjectivity and sentiment analysis of the Arabic language and dialects.And [5] described Ar-SenL a large scale Standard Arabic sentiment and opinion-mining lexicon using a combination of English SentiWordnet and Arabic WordNet. However, these resources lack detailed information necessary for efficient identification of fine-grained emotions in Arabic texts in general and social content in particular.

For example, given the Arabic tweets in Table 1, in addition to being able to identify the negative sentiments in tweets (1) and (2), and the positive sentiments in tweets (3) and (4), we want also to be able to recognize emotions of strong anger, fear, extreme happiness and joy in these tweets, respectively. To do this, however, requires linguistic resources that provide such fine-grained emotion information as well as emotional intensity scores for Arabic words and phrases.

TABLE 1
EXAMPLE ARABIC TWEETS EXPRESSING DIFFERENT EMOTIONS

1. @ahmadesseily	متغاظ ان كل الناس دول مش شايفين انه لا يصلح وزير حتّى؛ بكل كلامه المعسول اللي مليان أخطاء اللي بيغيّره لما يحتاج؛ متغاظ، أنا حر أتغاظ! *"I am furious that all those people don't see that he is not qualified to be a minister, despite his sugar-coated language that is full of mistakes that he changes whenever he needs. I am furious, I am free to getfurious!"*
2. @HadeerMostapha	طب انا مرعوبة من النتيجة يا اخوانا فعلا بكل ما تعنيه الكلمة *"Ok, bros, I am really scared in every sense of the word about the exam results"*

3.@Mr_atia_	يا أخي والله من ساعة ما شوفت الخبر ده وانا فرحان مزقطط أوي ... مشروع منتج @omhabib50 @HelpEGY هايل لناس غلابة تستاهل *"Bro, I swear to God I have been very happy and euphoric since I have heard the news ... It is an excellent project for the needy who really deserve it"*
4. @omdakhaaleed	وعهد الله طاير من الفرحة *"I swear to God, I am extremely delighted"*

Motivated by the lack of such a resource, this paper presents MASHAEIR (an Arabic word that means '*emotions'*), a fine-grained multi-dialect emotion thesaurus for Arabic. The construction of the thesaurus followed a corpus-based bootstrapping approach leveraging big data from Twitter. First, few Arabic adjectives expressing different emotions were used as search terms on Twitter from January 2007 to July 2015. Iterative bootstrapping was used to grow the seed and acquire new emotion-expressing adjectives. Second, concurrenceassociation measures and a longest-word-sequence identification method were used to retrieve multi-word expressions of emotions. Third, the final list of expressions was manually classified using well-recognized human emotions established by previous research in psychology that were adopted by prior work on fine-grained emotion identification. Fourth, coordination information was used to compute the emotional intensity indicated by these words and phrases. Finally, the location information in the profiles of the Twitter users in the corpus was used to establish Arabic dialectal preferences in the expression of emotions.

The remainder of this paper is organized as follows. Section 2 reviews previous work on the taxonomy and identification of fine-grained emotions. Section 3 details the different phases of bootstrapping MASHAEIR, describes the identification of the dialectal preferences in the expression of emotions, and presents a new method for computing the intensity of emotions expressed by the linguistic units in the thesaurus. Finally, Section 4 discusses the results and pertaining issues, and motivates future work.

2 RELATED WORK

Previous work on identifying fine-grained emotions in texts used well-recognized human emotions that have been established by psychological research (e.g. [17], [25] and [13]). Table 2 showsbasic human emotions as cited in[29].

TABLE 2
BASIC EMOTIONS IDENTIFIED BY EKMAN, IZARD AND PLUTCHIK

Ekman	**Izard**	**Plutchik**
Anger	Anger	Anger
Disgust	Contempt	Anticipation
Fear	Disgust	Disgust
Happiness	Fear	Fear
Sadness	Guilt	Joy
Surprise	Interest	Sadness
	Joy	Surprise
	Sadness	Trust
	Shame	
	Surprise	

These basic emotions provide the foundation for most of the recent emotion detection methods and ontologies, sometimes with modifications and additionsin order to cover emotions that are common on social media and were not included in the previous lists of basic emotions.

For example, [3] used Ekman's six basic emotion categories (i.e., *anger*, *disgust*, *fear*, *joy*, *love*, *sadness*, and *surprise*) to create a fine-grained emotion lexicon for English consisting of words that were automatically extracted from Roget's Thesaurus (1852). This lexicon was then used in automatically assigning an emotion label to each sentence in the given dataset, indicating the predominant emotion type expressed in the sentence.[26]presentedworkon annotating and detecting fine-grained emotions in English Twitter short messages, using Ekman's six basic emotions, in addition to *love* which they believed would be commonly found in informal text such as Twitter.[29] introduced EMOTIVE, afine-grained emotion ontology for English that covers the Plutick's eight basic emotionsexcept for '*anticipation*' and '*trust*', which were replaced with

'*confusion*' and'*shame*'. The authors' justification for this replacement was based on the frequent expression of these emotions with their differing levels of strengths as reactions to a varied range of events on Twitter.[15]alsoused Plutchik's eight basic emotions with no modifications to detect fine-grained emotions in Twitter messages.

In addition to their application to social media content, different versions of these basic emotions have also been applied to identify fine-grained emotions in various domains such as novels as in [19], gender differences in the expression of emotions in e-mail as in [21], news headlines as in [28], suicide notes as in[24], consumer views towards a product as in [22], [8]), and the stock market as in [9].

3 THESAURUS CONSTRUCTION

This section describes the different steps involved in the process: (1) iterative bootstrapping a list of adjectives, nouns, verbs and multi-words expressions of emotions, (2) measuring emotional intensity, and (3) identifying dialect preferences in emotion expression.

The construction of the proposed thesaurus followed a corpus-based approach leveraging big data from Arabic Twitter.Whenever applied in the following steps, the processes of manual filtering and classification were carried out by three Egyptian graduate students majoring in linguistics.Adopting a certain decision was based on the agreement of at least two out of the three. In the cases where dialects other than Egyptian Arabic were involved, the online Arabic dialect dictionary http://ar.mo3jam.com was frequently consulted(during the month of July 2015). In some cases, the author solicited dialect judgments from Twitter users in a number of Arab Countries.

To decide the basic emotions that would work better for the Arabic data, some preliminary tests were made with a sample data. It was found that a combination of the eight basic emotions in [29]and the emotions of *'guilt'*and *'interest'*from[17]seemed to provide a wider coverage for the Arabicsocial data. Given this, the basic emotions that were used included the following ten:*anger*, *disgust*, *fear*, *guilt*, *interest*, *happiness*, *sadness*, *shame*,*confusion*and *surprise.*

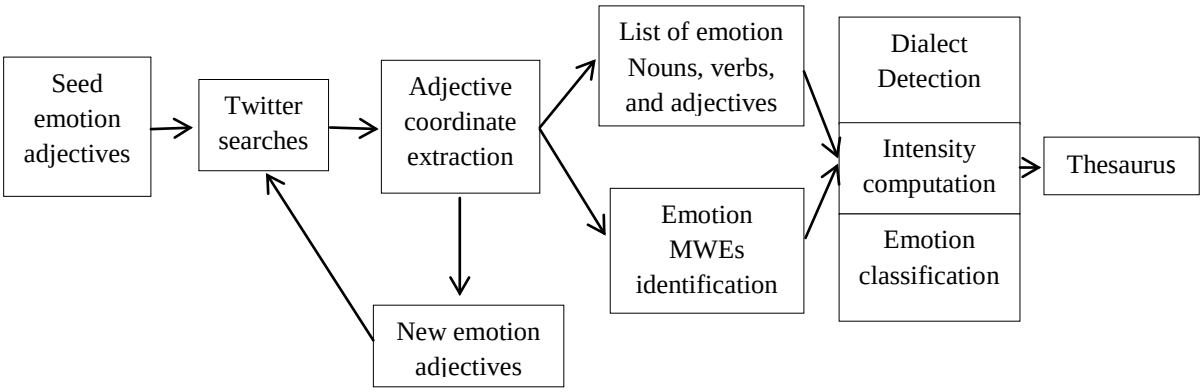

Figure 1: Methodology used in the construction of the proposed thesaurus

A. Bootstrapping Emotion Terms

Bootstrapping MASHAEIR was initialized using a seed of 10 Arabic adjectives expressing different emotional states, as given in Table 3. These adjectives were chosen such that: (1) they cover as many different emotions as possible, (2) half of them express positive emotions and the other half negative ones, (3) they are used frequently on Twitter, and (4) they are not specific to any dialect. The feminine and masculine forms of these adjectives were then used as Twitter search terms, once with the conjunctions '*و*/*and*' and '*أو*/*or*' and once without. The time span of the searches covered the period from January 2007 to July 2015. These date-based searches were automatically conducted to retrieve Arabic tweets on every day in the period mentioned. This time span was used to capture possible changes in linguistic expression of emotions over the years.

TABLE 3
SEED EMOTION ADJECTIVES[1]

حزين Hzyn sad	فرحان frHAn happy	خايف xAyf afraid	مرتاح mrtAH relaxed	متشائم mt$A}m pessimistic	متفائل mtfA}l optimistic	حيران HyrAn confused	متحمس mtHms eager	مصدوم mSdwm shocked	فخور fxwr proud

The adjectives coordinating with the seed adjectives were extracted, manually filtered to exclude spurious cases, and then used as the new bootstrapping seed. With every seed, the date-based Twitter searches were conducted in the same fashion described above. This bootstrapping process was applied iteratively until no more adjectives were acquired. In some cases, the elongated forms of some adjectives were retrieved and those were normalized to their original forms (e.g., *مبسووووط -> مبسوط/happy*).The output of this phase was 318 adjectives, including a large number of adjectives from different dialects of Arabic.Table 4 shows examples of the most and least frequent emotion adjectives that were identified, other than the seed adjectives.

TABLE 4
EXAMPLE MOST FREQUENT AND LEAST FREQUENT ADJECTIVES

Adjective	**Frequency**	**Adjective**	**Frequency**
زهقان zhqAn Bored	24,792	متوجع mtwjE agonized	491
متضايق mtDAyq upset	24,538	مغموت mgmwt upset	467
مبسوط mbswT happy	21,216	متضجر mtDjr annoyed	430
مخنوق mxnwq distressed	15,881	قانط qAnT despondent	425
منشكح mn$kH delighted	5,417	ملعلع mlElE euphoric	11

A list of Arabic nouns and verbs expressing emotions was then compiled in two different steps. The first step involved deriving the nouns and verbs corresponding to the final list of emotion adjectives.In the second step, new emotion nouns were identified using as Twitter search terms words and constructions that tend to be followed by emotions nouns: يحس/yHs/*feel*,يشعر/y$Er/*feel* and 'في حالة من'/ 'fyHAlpmn'/*'in a state of'*.The different morphological variations of these units generated 47 different forms that were used as Twitter search terms in the same date-based manner used previously with adjectives.A list was made of the nouns following these forms in the search results was compiled and filtered manually to identify only nouns that clearly expressed emotion. The output of these two steps consisted of 506 nouns. These nouns and their corresponding verbs, whenever available, were added to the list of emotion words comprising the thesaurus. This list contained nouns such as العار/AlEAr/*disgrace*, الوحشة/AlwH$p/*forlornness*, الغبطة/AlgbTp/*bliss*and العرفان//*gratitude*.

B. Emotion Multi-Word Expressions

Multi-word expressions (MWEs) are sequences of words that tend to co-occur more frequently than chance and are either idiosyncratic or decomposable into multiple simple words as in [6]. This term is flexibly used here to refer to any sequence of words that occurs frequently in the corpus, with no predefined linguistic criteria that sequence should satisfy.

[1]All transliterations were done using Buckwalter's Transliteration scheme.
URL: http://www.qamus.org/transliteration.htm

Possible MWEs indicating emotions were extracted from the corpus through identifying the longest sequences of words coordinating at least twice with two or more of theemotion-expressing adjectives identified in the previous phase. These sequences were then manually filtered to exclude instances that did not express emotions. The sequences identified were then used as Twitter search terms with and without the conjunction '*و/and*' and new instances of emotion MWEs were retrieved and manually filtered in the same fashion. This process was repeated until no more MWEs were identified. This resulted in 89such sequences that included collocations, idioms, and phrases,among other types of MWEs,as illustrated by the examples in Table 5.

TABLE5
EXAMPLE EMOTION MULTI-WORD EXPRESSIONS

مش طايق (نفسي – حد) I can't stand (myself, anybody)	طالع ديني I am fucked up
حالتي حالة feeling miserable	عال العال the best it can be
مقبل على الحياة feeling good about life	ضايقة بي الدنيا despondent
بالي مرتاح care-free	طاير في السما jubilant
مش ناقص I've had enough	مليش نفس أعمل حاجة I don't feel doing anything
قلبي مقبوض feeling apprehensive	راكبني مليون عفريت infuriated

C. Dialect Emotion Preferences

There are different groupings of the dialects of the Arabic language, based on the features used to measure differences and similarities.Common among these are the geo-linguisticand the socio-linguistic divisions of these dialects. Most of the research in the automatic Arabic dialect identification follows the geo-linguistic grouping (see for example,[33], [14], [7], [2] and [30]). According to this grouping, there are major five regional dialects of Arabic: Gulf, Iraqi, Levantine, Egyptian, and Maghrebi as in [7]. Sometimes, Sudanese, Mauritanian and Yemeni are considered threedifferent dialect classes of their own.This paper follows this form of dialect division, which is detailed as follows:

- **Gulf Arabic**: includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman.
- **Iraqi Arabic**: is the dialect of Iraq.
- **Levantine Arabic**: includes the dialects of Lebanon, Syria, Jordan, and Palestine.
- **Egyptian Arabic**: is the dialect of Egypt.
- **Maghrebi Arabic**: covers the dialects of Morocco, Algeria, Tunisia, and Libya.
- **Mauritanian Arabic**: is the dialect of Mauritania.
- **Yemeni Arabic**: is the dialect of Yemen.
- **Sudanese Arabic** :is the dialect of Sudan.

To acquire the dialect preference information for the identified emotion words and phrases, a list was made of the authors of the tweets in the corpus. The Twitter account profiles of those users were then automatically retrieved and searched for country and location information. Information on dialect preferences was only limited to users mentioning, either in Arabic or English, an Arab country or city in their profiles. Accordingly, a list of the English and Arabic names of the Arab countries and big cities was compiled, and a location was assigned to every account containing any of these names.

There were few cases where some profiles contained more than one country names. In these cases, the account was assigned to the country that was mentioned first in the account profile. In other cases, some users chose to use '*Gulf*' or '*Levant*'to refer to their location (66 and 33 users, respectively). These two location types were added to the list of the Arab countries/regions and were assigned to the corresponding profiles accordingly.

The total number of the Twitter accounts in the corpus was 170,060, and the total number of users that were identified to have Arab country location in the profile was 60,896, which constituted about 0.358 of the overall number of accounts.Table 6 shows the overall corpus size and the number of users in the corpus belonging to different Arab countries.

TABLE 6
DIALECT/COUNTRY DISTRIBUTION OF TWITTER USERS IN THE CORPUS

Dialect/Country	# of Users	Dialect/Country	# of Users	Country	# of Users
GULF (KSA)	22,328	**GULF** (Bahrain)	1,043	**SUD** (Sudan)	413
EGY(Egypt)	14,333	**IRAQI** (Iraq)	821	**MAGH** (Morocco)	303
GULF (Kuwait)	10,859	**LEV** (Syria)	746	**MAGH** (Algeria)	267
GULF (UAE)	3,298	**GULF** (Oman)	656	**MAGH** (Tunisia)	231
GULF (Qatar)	1,452	**LEV** (Lebanon)	552	**MAU** (Mauritania)	16
LEV (Jordan)	1,449	**YEM** (Yemen)	464		
LEV(Palestine)	1,097	**MAGH** (Libya)	461		

Total number of tweets in the corpus: 8,341,836
Total number of Twitter users in the corpus: 170,060
Number of users in the corpus with profile Arab country information: 60,896 (0.358 of total users)

The process of deciding the dialectal preference for a certain emotion termwasmainly based on the frequency of this term use among the Twitter users from the countries where this dialect is dominant.The original idea was to use a set of association measures to test the dialectal preferencesfor a given emotion term. However, the sample corpus is not, by any means, representative of the different dialects of Arabic given the different penetration rates of Twitter users in different Arab countries.This sampling effect was clear in the strong biases that showed up in the results of some experimental tests using a number of association measures. Accordingly, the decision was to use the absolute frequencies of the emotion terms to establish their dialectness in this preliminary version of the thesaurus, leaving a more accurate method for future versions using a larger and more representative dialect corpus.

Given this, the dialectness of emotion terms was established and encoded using the two following procedures.First, a term should be used by at least 10 Twitter users in a given dialect to be considered part of this dialect. This arbitrary threshold was meant to exclude cases where users from a given dialect were commenting on some emotion terms used by other dialect users. Using this threshold,a term used by users from all dialects was considered non-dialectal, and in this case, the dialect preference for this term was annotated as *NON-DIALECT* in the thesaurus.In other cases, the dialect preferences for a given term were encoded in a descending order based on its frequency of use in a given dialect. Table 7 shows example terms and their dialect preferences.

TABLE 7
SAMPLE TERMS AND THEIR DIALECT PREFERENCES

Term	Dialect Preferences
خائف, سعيد, متضايق, مصدوم, جبان, حائر, حزين	*NON-DIALECT*
منفسن	*EGY, GULF, LEV*
مبضون	*EGY*
متعفس	*GULF*

It is important to mention in this context that there were some cases where a given emotion term had different senses and expressed different emotions in different dialects. For example, the adjective 'طفشان' is frequently used in the Levantine and Gulf dialects to mean 'feeling bored', in addition to the other sense in the context 'طفشان من البيت'to mean 'leaving the house due to anger', predominantly in the Egyptian dialect.In these cases, dialect judgments were solicited from a number of Twitter users that were picked randomly from the user list mentioned above. This and other related issues are not elaborated upon here are left for future research.

D. Emotion Classification

Emotion words and phrases the final list of were then classified manually according to two parameters. The first was a semantic polarity-based classification, i.e. negative and positive. The other was an emotion-based classification into the ten basic emotions mentioned above, i.e., *anger, disgust, fear, guilt, interest, happiness, sadness, shame, confusion* and *surprise.* The annotation task was carried out by three Egyptian graduate students majoring in linguistics in the fashion previously stated.

During the annotation processes, there arose some issues related to the cultural perspective regarding the polarity and classification of some emotion terms/concepts. For example, terms derived from the concept of خشوع/'state of being God fearing' should be classified under '*fear*', and consequently should have a negative polarity. However, this concept has a predominantly positive connotation in the Arab/Islamic culture in general. In this case, the ad-hoc solution was to assign a positive polarity to terms though they are related to '*fear*', an emotion that normally has a negative semantic polarity. The same thing applied to other concepts such as 'خجل, حياء/shyness'.(The cultural 'sensitivities' in emotion polarity and classification seem to have significant effects on emotion identification in texts and deserve serious future research). Table 8 shows some example terms and their final emotion classification.

TABLE8
SAMPLE BASIC-EMOTION TERMS

Emotion	**Examples**
Anger	خلقي ضايق ,باتخانق مع دبان وشي,غضبان,....
Disgust	ممتعض ,مبضون ,قرفان ,طفشان ,زهقان,...
Fear	قلبي مقبوض ,خاشع ,مذعور ,مفزوع ,مرعوب ,خايف,...
Guilt	ضميري بيأنبني ,مقصر في حق ,غلطان,حاسس بالذنب,...
Interest	داخل دماغي ,معجب ,متلهف ,مشتهي ,مشتاق ,شغوف,...
Happiness	مزاجي عال العال ,مزقطط ,مبسوط ,فرحان,...
Sadness	مكتئب ,تعيس ,مقفلة معايا ,حزين ,زعلان,...
Shame	متهان ,مكسوف من نفسي ,خائب,...
Confusion	مشتت ,مش فاهم حاجة ,ضايع ,حيران ,تايه,...
Surprise	مش مصدق نفسي ,مندهش ,مستغرب ,مصدوم,...

E. Measuring Emotional Intensity of Terms

Intensity score is the result of separate intensity scores that quantify the following two observations about the coordination properties of emotion terms. The first observation is that the terms expressing more tense emotions generally tend to conjoin more frequently with other terms expressing emotion of the same polarity and less with the terms of the opposite polarity. This idea was used by [16]to identify the semantic polarity of adjectives in English. This observation is used to compute the polarity-based emotion intensity of emotion terms,$EI(y_p)_{pol}$, as follows:

$$EI(y_p)_{pol} = \frac{\sum_{i=1}^{i=n} frequency(y_p, x_{pi})}{frequency(y_p)} \quad (1)$$

where the nominator is the number of times term y of semantic polarity p conjoins with other terms, $x_1 \ldots x_n$, of the same polarity, and the denominator is the number of times term y occurs in coordinate structures in the corpus.

The other observation is that terms expressing more intense emotions tend to follow terms expressing less intense emotions in a coordinate structure. This observation is to calculate the position-based emotion intensity of emotion terms,$EI(y)_{emo}$, as follows:

$$EI(y)_{emo} = \frac{\sum_{j=2}^{j=l} frequency(y)_j}{frequency(y)} \quad (2)$$

where the nominator is the frequency of occurrence of the given term in a non-initial position in a coordinate structure (*2,3...l*), where *j* is the position of the term *y* in the coordinate structure, *l* is the length of this structure as measured by the number of words occurring in the structure, and the denominator is the number of times term *y* occurs in coordinate structures in the corpus.

The final emotional intensity score for a given term in the thesaurus is simply the average of the two scores in Eq.1 and Eq. 2. This score is a value larger than zero and less than or equal to 1, where 1 indicates maximum emotional intensity. Table 9 shows example emotion adjectives and their corresponding emotional intensity scores.

TABLE 9
EXAMPLE EMOTION ADJECTIVES AND THEIR INTENSITY SCORES

Term	**Polarity**	**Emotion**	**Intensity**
مشتهي/m$thy/craving	Positive	Interest	0.84
مرعوب/mrEwb/scared	Negative	Fear	0.77
متغاظ/mtgAZ/furious	Negative	Anger	0.68
مزقطط/mzqTT/euphoric	Positive	Happiness	0.68
...			
حزين/Hzyn/sad	Negative	Sadness	0.52
سعيد/sEyd/happy	Positive	Happiness	0.52

4 THESAURUSCOVERAGE EXPERIMENT

The coverage is meant to roughly measure how much the proposed thesaurus captures the words and phrases used by the users of Arabic Twitter to express their fine-grained emotions. In order to do this, a test Twitter corpus was built and used as follows:

a. A list of hashtags was made of the adjectives in the thesaurus.
b. Every hashtag was then used as a Twitter search term. The time span of the searches was the month of July 2015, which was not part of the Twitter corpus used in constructing the thesaurus.
c. All tweets that contained only hashtags, and retweeted tweets were removed.
d. In the tweets that remained, the emotion hashtag terms were removed.

The resulting test corpus contained 21,816 unique tweets.The coverage was then tested in two different ways: The first with no preprocessing of the tweets in the test corpus, and the other with light preprocessing. To carry out this preprocessing, a small script was written to remove the coordinating conjunction '*و*/waw/*and*', and the prepositions and pronouns attached to the target words. Table 10 shows the results. Without preprocessing, the result was that only 7,145 tweets contained emotion terms, yielding a coverage rate of almost 0.33.Applying preprocessing to the test corpus resulted in a 0.02 increase in the coverage rate.

As for the coverage in terms of the ten fine-grained emotions adopted in the proposed thesaurus, all these emotions were covered with different degrees. As Figure X shows, the emotions of *Anger*, *Happiness* and *Sadness*were the top most covered emotions in the test corpus.It is also clear that the coverage the thesaurus achieves with a given emotioncorrelates with the number of words and phrases that express this emotion in the thesaurus.After examining a sample of the test tweets that were not covered by the thesaurus, it was observed

TABLE 10
THESAURUS COVERAGE RATES WITH AND WITHOUT PREPROCESSING

Coverage without preprocessing	Coverage with preprocessing
0.33	0.35

that the coverage of the emotion MWEs was very poor, compared to the wide coverage of nouns and adjectives. This result calls for further research to enrich the proposed thesaurus with such expressions.

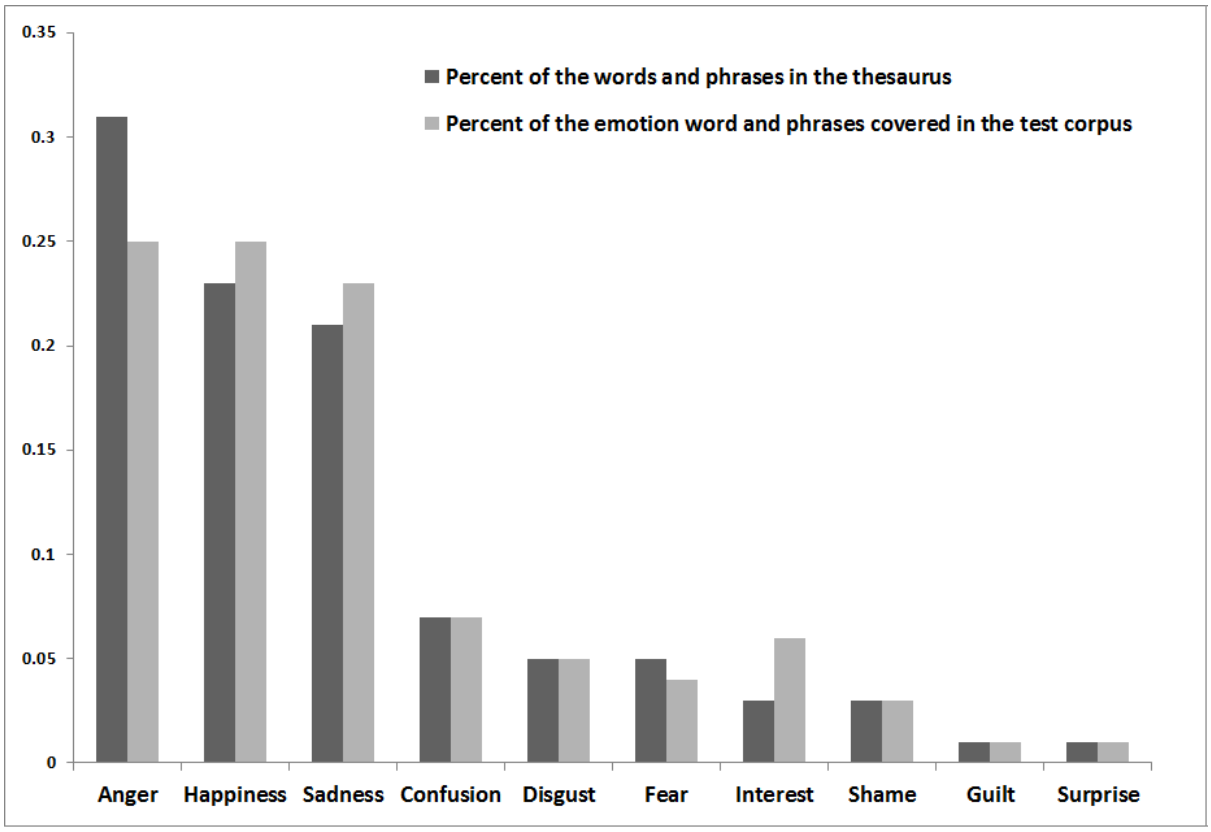

Figure 2: Thesaurus coverage of emotions in the test corpus

5 CONCLUSION AND FUTURE WORK

This paper has presented the steps for constructing a fine-grained multi-dialect thesaurus for Arabic leveraging a Twitter corpus from 2007 until July 2015. The paper has suggested a lexicon of basic emotions for Arabic, as well as a new method for measuring the intensity of the emotion terms. A simple coverage test was also presented briefly. During the process of constructing this thesaurus some important issues arose regarding some language- and culture-specific aspects of emotion taxonomies as well as dialect preferences. Another issue that also arose and was not discussed in the paper was gender differences in the expression of emotions. This thesaurus provides a starting point for future efforts for building a linguistic resource for the identification of fine-grained emotions in Arabic (social) texts, where dialect and gender issues will be examined more elaborately.

ACKNOWLEGEMENS

The author is grateful to Sara Awad, EsraaAbdelzaher, and NermeenYousry for their valuable help in the tasks of manual filtering and classification of parts of the data used in this paper. Appreciation is also due to the Twitter users who provided dialect judgments concerning some words and expressions. The usual disclaimers apply.

REFERENCES

[1] Abdul-Mageed, M., & Diab, M. Sana: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May 26-31, 2014.

[2] Alorifi, Fawzi S. *Automatic Identification of Arabic Dialects Using Hidden Markov Models*. ProQuest, 2008.

[3] Aman, S., &Szpakowicz, S. Using Roget's Thesaurus for Fine-grained Emotion Recognition. In *IJCNLP*, 2008*:* 312-318.

[4] Aman, S., and Szpakowicz, S. "Identifying expressions of emotion in text." *Text, Speech and Dialogue*. Springer Berlin, 2007.

[5] Badaro, Gilbert, et al. "A large scale Arabic sentiment lexicon for Arabic opinion mining." *ANLP 2014*:165-173.

[6] Baldwin, Timothy. "Compositionality and multiword expressions: Six of one, half a dozen of the other." *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July*. 2006.

[7] Biadsy, Fadi, Julia Hirschberg, and Nizar Habash, "Spoken Arabic dialect identification using phonotactic modeling."*Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. Association for Computational Linguistics, 2009:53-61, Athens, Greece.

[8] Bollen, Johan, Huina Mao, and Alberto Pepe. Modeling Public Mood and Emotion: Twitter Sentimentand Socio-Economic Phenomena. In Proceedings of the *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 17-21 July 2011, Barcelona, Spain.

[9] Bollen, Joha, Muina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011b,1–8.

[10] Brynielsson, Joel, et al. "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises." *Security Informatics* 3.1 (2014): 1-11.

[11] Buitinck, Lars, et al. "Multi-emotion Detection in User-Generated Reviews." *Advances in Information Retrieval*, Springer International Publishing, 2015. 43-48.

[12] Counts, Munmun De Choudhury Scott, and Michael Gamon. "Not all moods are created equal! Exploring human emotional states in social media." *in Proceedings of ICWSM*, 2012.

[13]Ekman, Paul. "An argument for basic emotions."*Cognition & emotion* 6.3-4 (1992): 169-200.

[14]Habash, Nizar Y. "Introduction to Arabic natural language processing." *Synthesis Lectures on Human Language Technologies* 3.1 (2010): 1-187.

[15]Hasan, Maryam, ElkeRundensteiner, and Emmanuel Agu. "Emotex: Detecting emotions in twitter messages." In *Social Com-Stanford*, 2014.

[16]Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives. "*Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics*.Association for Computational Linguistics, 1997.

[17]Izard, Carroll E. "Emotion theory and research: Highlights, unanswered questions, and emerging issues." *Annual review of psychology* 60 (2009): 1.

[18]Jennifer, D. "Affective Text based Emotion Mining in Social Media." *International Journal of Advanced Research in Computer Science and Management Studies*, 2:3, March, 2014:86-94.

[19]Mohammad, Saif M. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage*, 2011: 105–114.

[20] Mohammad, Saif M., and Svetlana Kiritchenko. "Using hashtags to capture fine emotion categories from tweets, "*Computational Intelligence* 31.2 (2015): 301-326.

[21] Mohammad, Saif M., and Tony Yang.Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and SentimentAnalysis (WASSA)*, 2011, pages 70–79.

[22] Pak, Alexander, and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In*Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.

[23] Paltoglou, Georgios. "Sentiment analysis in social media. "*Online Collective Action*. Springer Vienna, 2014.3-17.

[24] Pestian, John P., Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin B. Cohen, John Hurdle, and Christopher Brew. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 2012

[25] Plutchik, R. The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 2001, *89*(4), 344-350.

[26] Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. EmpaTweet: Annotating and Detecting Emotions on Twitter. In *LREC*, 2012, (pp. 3806-3813).

[27] Sintsova, Valentina, Claudiu Musat, and Pearl Pu. "Semi-Supervised Method for Multi-category Emotion Recognition in Tweets. "Data Mining Workshop (ICDMW), 2014 IEEE International Conference on.IEEE, 2014.

[28] Strapparava, Carlo, and Rada Mihalcea. Learning to Identify Emotions in Text, In *Proceedings of the ACM Conference on Applied Computing*, 2008.

[29]Sykora, M. D., Jackson, T. W., O'Brien, A., & Elayan, S. Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *Computer Science and Information Systems Journal*, 2013.

[30] Versteegh, Kees. *The arabic language*. Edinburgh University Press, 2014.

[31]Wang, Wenbo, et al. "Harnessing twitter" big data" for automatic emotion identification." Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012.

[32]Yan, Jasy Liew Suet. "Expanding the Range of Automatic Emotion Detection in Microblogging Text. "*EACL* 2014, 2014: 38.

[33]Zaidan, Omar F., and Chris Callison-Burch. "Arabic dialect identification."*Computational Linguistics* 40.1 (2014): 171-202.

BIOGRAPHY

Dr.Khaled Elghamry is an Associate Professor of (Computational) Linguistics, Faculty of Alsun, Ain Shams University, Egypt. Dr. Elghamry obtained a Ph.D. in Computational Linguistics, 2004, Indiana University, USA. He was a visiting scholar, Department of Linguistics, University of Florida 2007-2010. Dr. Elghamry is the Co-Founder of the Midwest Computational Linguistics Forum, Indiana University, and also the Co-Founder of the Arabic Digital Content Statistical Database, and the Arabic Digital Content Foundation Report. Dr. Elghamry presented and published research in international conferences and journals on different technical issues in the automated processing of the Arabic language and content, Web and textmining, sentiment analysis, and tracking of online public opinion.

مشاعر: إنشاء مكنز لغوي متعدد اللهجات للمشاعر الدقيقة للغة العربية باستخدام تويتر

خالد الغمري
قسم اللغة الإنجليزية – كلية الألسن – جامعة عين شمس

خلاصة

يعدالمحتوى المتاح على مواقع التواصل الاجتماعي مثل تويتروفايسبوك مصدرا ثريا لمشاعرمستخدميها إزاء المنتجات والقضايا والأحداث الجارية. وعليه فقد تزايد الاهتمام البحثي في السنوات الأخيرة بتصنيفات تفصيلية لهذه المشاعر تتجاوز تصنيفها البسيط إلى مشاعرسلبية وإيجابية

فقط. وفي ضوء ذلك تقدم هذه الورقة المكنزاللغوي "مشاعر" و هومورد معجمي للتعامل مع تصنيفات المشاعرفي النص العربي مع وضع اختلاف اللهجات العربية في تعبيرهاعن هذه المشاعرفي الاعتبار. يعتمدإنشاءالمكنزعلى عددضخم من الرسائل القصيرة بالعربية من موقع التواصل الاجتماعي "تويتر" في الفترة من يناير 2007 حتى يوليو 2015. والمكنزيضم مفردات وعبارات عربية تستخدم في التعبيرعن المشاعرالدقيقة وكذلك التفضيلات اللغوية في اللهجات العربية المختلفة في هذا السياق. كما تقدم الدراسة طريقة جديدة لقياس قوة المشاعرالتي تعبرعنها هذه المفردات والعبارات. ولقياس قدرة هذا المكنز للتعامل مع المحتوى الاجتماعي العربي، تقدم الدراسة تقييما سريعا وبسيطا باستخدام عينة من رسائل تويتر باللغة العربية. والقصد من هذا المكنز أن يكون نواة لمورد معجمي عربي للمشاعرالمختلفة قابل للتحديث والتطوير في المستقبل.