A Dataset to Support Sexist Content Detection in Arabic Text

Oumayma El Ansari^(✉), Zahir Jihad, and Mousannif Hajar

LISI Laboratory, Faculty of Sciences Semlalia, Cadi Ayyad University,
Marrakech, Morocco

ansari.oumaima@gmail.com,
{j.zahir,mousannif}@uca.ac.ma

Abstract. Social media have become a viral source of information. This huge amount of data offers an opportunity to study the feelings and opinions of the crowds toward any subject using Sentiment Analysis, which is a struggling area for Arabic Language. In this research, we present our approach to build a thematic training set by combining manual and automatic annotation of Arabic texts addressing Discrimination and Violence Against Women.

Keywords: Sentiment analysis · Arabic language

1 Introduction

Violence Against Women (VAW) is one of the most commonly occurring human rights violations in the world, and Arab region is no exception. In fact, UNWOMEN [3] reports that 37% of Arab women have experienced some form of violence in their lifetime. It's been demonstrated that discriminatory attitudes are still posing a challenge for women's status in the Arab States.

There is a variety of methods to measure attitudes and opinions towards a subject. Data generated by Internet activity, especially Social Media activity where users have the freedom of speech, is an interesting data source that can be used to evaluate the public opinion regarding both Discrimination (DAW) and Violence Against Women (VAW).

Sentiment Analysis uses data, typically from Social Media, to analyze the crowds' feelings toward a certain subject. It's the task of determining from a text if the author is in favor off, against or neutral toward a proposition or target [4]. In other words, it classifies a subjective text into different polarities: Positive (e.g. it's the best phone ever!), Negative (e.g. it sucks!) and Neutral (e.g. the new version is out) [5].

There are two main approaches to Sentiment Analysis: Machine learning (ML) based and lexicon-based. ML methods consist of using pre-annotated datasets to train classifiers. In lexicon-based methods, the polarity of a text is derived from the sentiment value of each single word in a given dictionary. ML based methods usually give higher accuracy [6], however these methods require a good quality training set to provide accurate and precise classifiers.

We are interested in applying Machine learning based Sentiment Analysis to the topic of ‘Violence and Discrimination against Women’. However, there is, as far as we know, no existing annotated dataset featuring Arabic texts addressing this topic.

In fact, building an annotated training set is a fastidious and time consuming task requiring the involvement of human annotators. In this research, we present our approach to build a thematic training set by combining manual and automatic annotation of Arabic texts addressing Discrimination and Violence against Women.

In this work, we make three main contributions:

- A. We develop an initial training set [7] that contains Arabic texts related to Discrimination and Violence Against Women and annotated by humans,
- B. We propose a method that automatically extends the initial training set. In fact, we use the initial training set to generate a list of key expressions, and use them to produce a new expanded training with roughly the same characteristics.
- C. We analyze the inheritance level of polarities from the Initial Dataset to the new collected data.

The remainder of this paper is organized as follows: Sect. 2 is a description of the general process. Section 3 describes the steps we followed to generate key expression. Section 4 represents the phase of building the new extended training set and the results of our work.

2 The Approach

In this approach, we develop an initial training set that contains Arabic texts related to Discrimination and Violence Against Women and annotated by human volunteers. Then, we use this initial training set to retrieve two lists (positive and negative) of key-expressions that represent the most significant terms used by Arab-speaking internet users to express themselves either negatively or positively towards Discrimination and Violence Against Women. Based on the generated lists we build a new training set on which we studied the inheritance of polarities (Fig. 1).

The general process is composed of two main phases: A. the automatic generation of key expressions. B. Building and analyzing the new training set.

Fig. 1. The general process

3 Generation of Key Expressions

3.1 Initial Dataset

The starting point of this research is an initial pre-annotated training set which consists of a sample of raw text data containing tweets and some YouTube comments. A number of tweets in the sample were collected during the International Women's Day in 2018. Human annotators had to choose between six labels to annotate the different pieces of text:

- Off-Topic: If the text doesn't address a topic related to women's and girls' rights or realities,
- Neutral: If the text presents neutral information,
- Positive: If the text provides a positive position or opinion,
- Negative: If the text provides a negative position or opinion,
- Mixed: If the text provides a mix of positive and negative positions or opinions.

The labels 'Neutral', 'Positive', 'Negative' and 'Mixed' apply only for text data that actually addresses a topic related to women's and girls' rights.

3.2 Preprocessing

Data Cleaning. First, we started by extracting separately the positive and negative comments from the dataset. We retrieved a smaller set of 518 comments: 292 positive and 226 negative.

The preprocessing phase is slightly delicate for morphological languages such as Arabic. We started with cleaning data from any noises in order to have a clean Arabic piece of texts ready to be processed instead of noisy comments. Below all the components that we eliminate from data:

a - Diacritics: These diacritics express the phonology of the language and, in contrary of English, Arabic words could be read with or without their diacritics. Example:

المرأة سياقة سياقة المرأة

b - Emojis: Users on social media tend to frequently use Emojis to express their opinions and feelings. In our case we removed all the emoticons and symbols from our set.

c - Punctuation: in the field of Natural Language Processing, the presence of punctuation affect directly the treatment of texts. To avoid bad results, we eliminated any presence of punctuation in our comments.

d - Numbers: We removed all the numerical digits.

Normalization and Stop Words Removal. Normalization is the process of transforming words to a standard format. In Arabic language, the phonetic sound of [i] [a] [u] could be written in multiple syntactic forms, it depends on its location in the word. We implemented a normalized format to this sounds to reduce conflicts during the treatment of the texts (Table 1).

Table 1. Normalized letters.

Letters	normalization
ا / آ / أ / إ /	ا
ى	ي
ئ / و	ء
ئ	ء

Stop words removal helps to eliminate unnecessary text information. We used a predefined set of a 3616 Arabic stop words to clean data. However, we faced a huge difficulty in this step because most of the time Arabic internet users tend to skip putting a space between a short stop word and the word that follows, the system consider it as one different word. Example: **و** is a stop word but it will not be removed as its attached to المرأة:

المرأة و والمرأة

3.3 Frequency Calculation

Now as we cleaned our data, we must prepare it for the next step that consists of finding key expressions based on frequencies calculation.

Tokenization is the task of chopping a text into pieces, called tokens. Tokens do not always refer to words or terms but it could represent any sequence of semantic expressions. Depending on the typology of our work, we've choose to fragment the text into words, then we calculated the frequency of all the expressions of one to five words (n-grams) in order to retrieve the most significant expressions toward our topic.

At this stage, and whether for negative or positive tokenized text, we retrieved the top 20 most frequent expressions for each n-gram ($n = 1 \rightarrow 5$). We have end up with 100 expressions for each polarity.

3.4 Final Key Expressions

The preprocessing phase that we carried out was not sufficient to have good result. In fact, the 200 expressions that we retrieved were not very convincing due to many parameters, in what follows the actions that we did to filter the expressions:

Remove “Religious” Expressions. Arab speakers tend usually to use general purpose religious expressions in their discourse, the same case for Arabic internet users, which explains the strong presence of such terms in the retrieved expressions. During pre-processing, we replace these religious expressions by the keyword “RE”.

Example:

حسبي الله, الرسول صلى الله عليه و سلم

Remove Named Entities Expressions. The collected data is usually influenced by major events that invade social networks during the time of scrapping which leads to a redundancy of a person's name or an organization, for example, in collected data:

الشيخ فوزان, الملك سلمان

Remove Insults. Insults are frequently used by net surfers especially in social media platforms. Example:

غبية

Remove Expressions Present in Both Negative and Positive Lists. In order to leave only significant expressions toward either positive and negative polarity, it's obvious to remove expressions present in both lists at once.

Normalizing Identic Expressions. This step is about conserving only one of various expressions that give the same meaning. For example both “قيادة” and “قيادة” are synonyms to the word “drive”:

سياقة المرأة قيادة المرأة

Final Lists. After filtering the 200 expressions, we kept 6 expressions for each polarity: RE = Religious Expression

Positive	Negative
ضرب المرأة 'Hit the woman'	فرن بيوتكن 'Stay home'
عقلية ذكورية 'Male mentality'	تطلق المرأة 'get divorced'
امسك بمعروف RE	تكشف وجهها 'Show her face'
تسريح بإحسان RE	تلتزم بالحجاب 'keep the Hijab'
ختان المرأة 'Female circumcision'	مطبعة 'Obedient women'
معاملة الرجل للمرأة 'Man treats woman'	تبرجن تبرج RE

4 Building the New Expanded Dataset

In phase A, we generated key-expressions, based on an initial dataset annotated by humans. Each polarity (positive and negative) is represented by 6 key expressions.

When we use these expressions as seeds to collect new data, we expect the collected data to have the same polarity as the seeds. In this section, we describe the data collection process and try to answer this question: Did the new data inherit the polarity of the key-expressions used to collect it?

4.1 Data Collection

Using Twitter Developer API, we collected tweets using the pre-generated list of key expressions. We finally retrieved 1172 tweets: 573 negative, 599 positive.

4.2 Evaluation

We assess the quality of the expanded dataset by direct human judgment. This consists of human volunteers reviewing and annotating the collected texts using one of four labels:

- Off-Topic: If the text doesn't relate to our topic,
- Positive: If the text represents a positive opinion toward women,
- Negative: If the text represents a negative opinion toward the topic,
- Neutral: For texts that describes a neutral information or opinion.

With this step, we were able to distinguish between true and false negative texts (i.e. positive texts).

4.3 Analysis

Data Collected with Positive Key Expressions. With these key expressions we retrieved good results, the majority of the collected tweets are true positive with a percentage as high as 86%. Negative tweets represent only 4% (Fig. 2).

Fig. 2. Results for data collected with positive keywords

Data Collected with Negative Key Expressions. In this case, the results were not satisfying. True negative tweets represented only 42% of the whole collected data, the proportion of neutral tweets was remarkably high with a percentage of 34%. In fact, a substantial number of the retrieved tweets in the negative set consisted of verses of Quran, which were annotated as Neutral (Fig. 3).

Fig. 3. Results for data collected with negative keywords

5 Conclusion

In this work, we presented an approach to build a training set by combining manual and automatic annotation of Arabic text. The preprocessing was very challenging due to the complexity of the language and the typology of data in social media.

The obtained results were of good quality: we built a final training set of 1690 entries. However, even if the approach gives excellent results for positive key-expressions, the inheritance of polarities must be improved for negative key-expressions which we will further investigate in future works.

References

1. <https://unstats.un.org/sdgs/indicators/indicators-list/>. Accessed Feb 2019
2. Vaitla, B., et al.: Big data and the well-being of women and girls: applications on the social scientific frontier (2017)
3. <http://arabstates.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures>. Accessed Feb 2019
4. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: A dataset for detecting stance in tweets. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3945–3952, May 2016
5. Abdul-Mageed, M., Diab, M.T.: AWATIF: a multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In: LREC, vol. 515, pp. 3907–3914, May 2012
6. Taboada, M.: Sentiment analysis: an overview from linguistics. *Ann. Rev. Linguist.* **2**, 325–347 (2016)
7. Zahir, J.: Mining the web for insights on violence against women in the MENA region and Arab states (2019)