

# An Empirical Evaluation of Arabic-specific Embeddings for Sentiment Analysis

Amira Barhoumi, Nathalie Camelin, Chafik Aloulou, Yannick  
Estève, Lamia Belguith

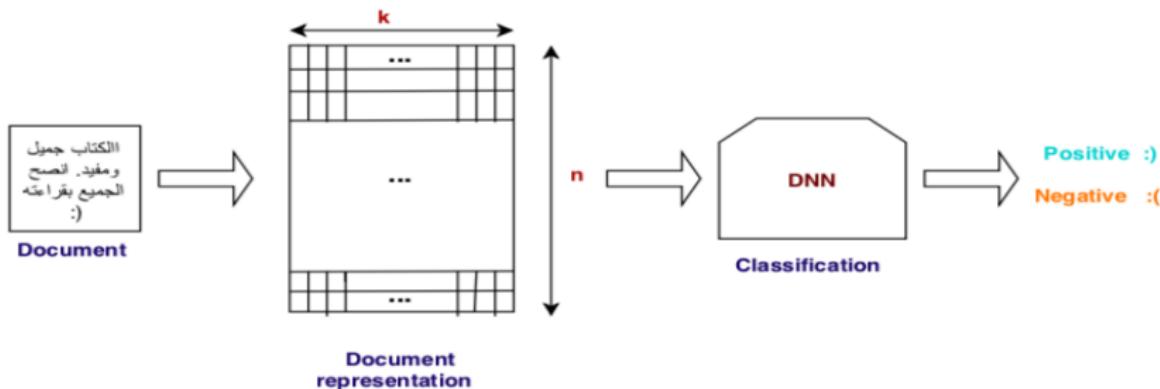
LIUM, Le Mans University, France

MIRACL, Sfax university, Tunisia



# Introduction

- Sentiment Analysis (SA) framework :
  - given a textual statement,
  - identify its polarity : positive or negative.



**$k$  = embedding dimension**

**$n$  = document length**

**DNN = Deep Neural Network**

## State of the art

- Recent works use Deep learning techniques :
  - based on Convolutional Neural Network (CNN) ([Dahou et al., 2016], [Alayba et al., 2018], [Dahou et al., 2019]).
  - based on Long Short-Term Memory (LSTM) ([Hassan, 2017], [Heikal et al., 2018], [Al-Smadi et al., 2018]).
- Network inputs = **word embeddings**.
- Words = space separator units.
- Do not take into account specificity of Arabic language (agglutination and morphological richness) in the embedding space.

# Specificity of Arabic language : Agglutination and morphological richness

- Agglutination
  - Phrases could be composed with only one word.  
 i.e. : أَعْجِبُكَ (Do you like it ?)
  - Word = inflected form + clitics (proclitics and enclitics).

| Word    | Translation       | Decomposition  |           |
|---------|-------------------|----------------|-----------|
|         |                   | Inflected form | Clitics   |
| اسيعجبه | will he like it ? | يَعْجِب        | أ + س + ه |
| فسيعجبك | you will like it  | يَعْجِب        | ك + س + ف |
| ويعجبهم | and they like it  |                | و + هم    |

- Morphological richness : root + schemes  
 i.e. : conjugation of يَعْجِبُونَ : عَجَبٌ ، يَعْجِبُانَ ، يَعْجِبُونَ

# Specificity of Arabic language : Agglutination and morphological richness

- 6 different lexical units :
  - word : space separator unit.
  - token : inflected form and clitics (Farasa tool<sup>1</sup>).
  - token\clitics :inflected form (Farasa tool).  
Clitics do not usually affect the polarity of words.
  - Lemma : canonic form (Farasa tool).
  - stem : root of word (Tashaphyne tool<sup>2</sup>).
  - light stem : stem + infixes (Arabic light stemmer<sup>3</sup>).
- Granularity of words could impact the embedding quality.

---

1. <http://qatsdemo.cloudapp.net/farasa/>

2. <https://pypi.org/project/Tashaphyne/>

3. <https://github.com/motazsaad/arabic-light-stemming-py>

# Embedding Models

- Word2vec (w2v) [Mikolov et al., 2013b] :
  - words sharing common context are closely located in the embedding space.

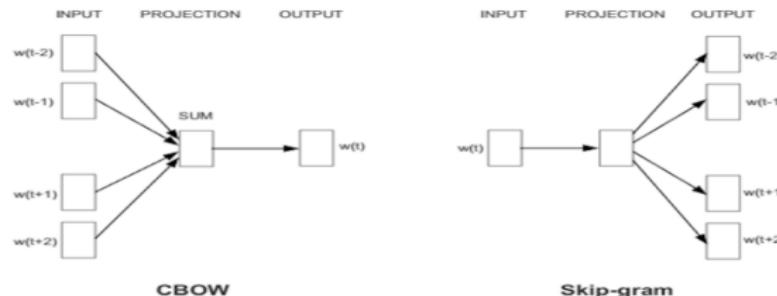


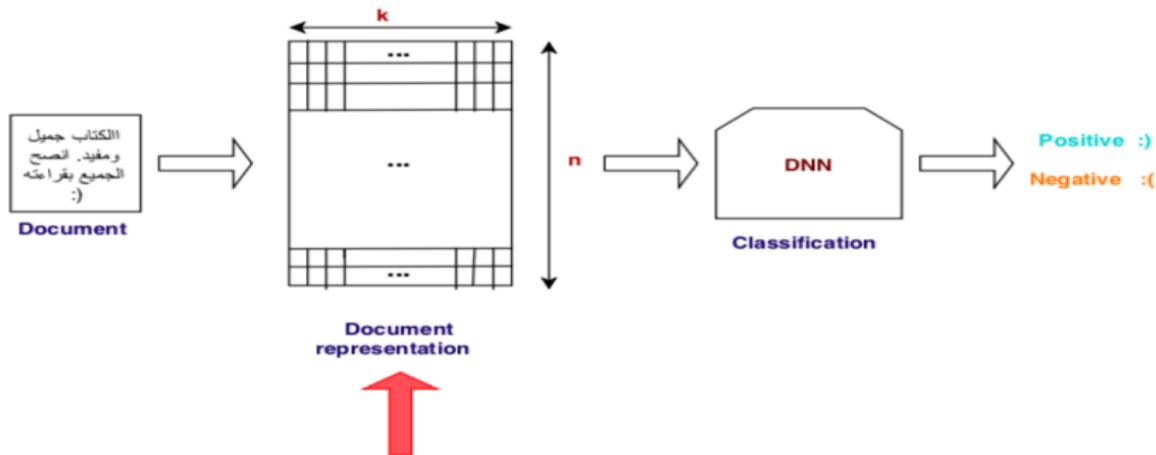
Figure – Skip-gram and CBOW architectures [Mikolov et al., 2013a].

- Skip-gram ≫ CBOW [Dahou et al., 2016, Barhoumi et al., 2018]

# Embedding Models

- **Fasttext (ft)** [Bojanowski et al., 2016] :
  - Extension of word2vec.
  - embeddings for unseen rare words.
  - sub-word (n-gram character) information.  
i.e. : for the word "where", add 2 symbols (< and >) <where>  
if n = 2, sub-words = {<w , wh, he, er, re, e>}

## Document representation



- **Document length  $n$  :**

- Hypothesis : length distribution follows Gaussian law.

$$n = \text{mean} + 2 \times \text{standard deviation} \quad (1)$$

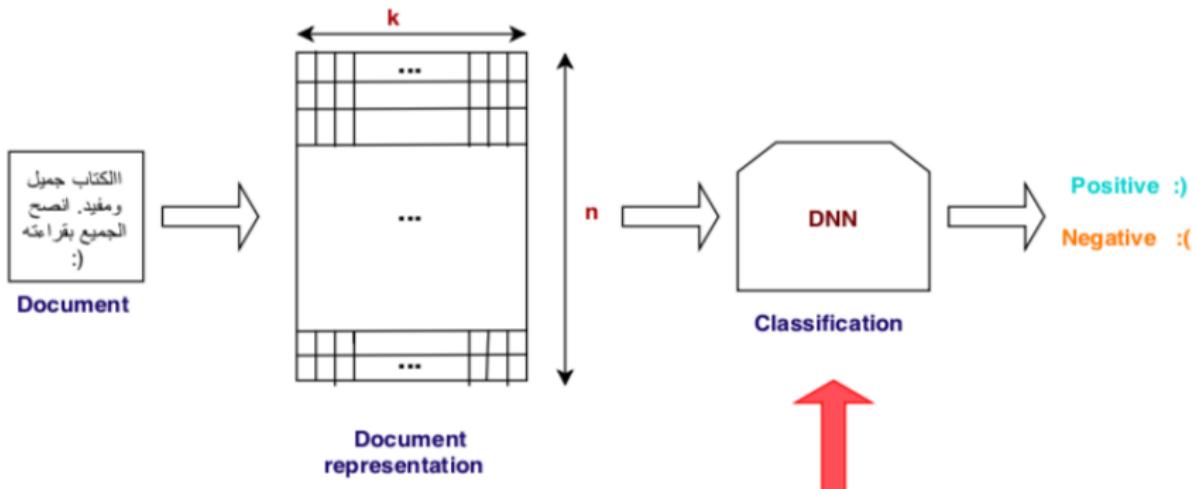
- Empirical rule : Probability of documents  $\approx 0.9545$

## Document representation

What type of padding/truncating (Begin, end, extremities) ?

- Protocol :
  - Splitting the document into 3 parts.
  - Analyse polar words and negation terms for each part.
  - Informativeness of each part :
    - First part : post padding/truncating.
    - Second part : padding/truncating on extremities.
    - Last part : pre padding/truncating

## DNN Component



- CNN-based system.
- BiLSTM-based system.

## CNN-based system

- Architecture similar to [Dahou et al., 2016].

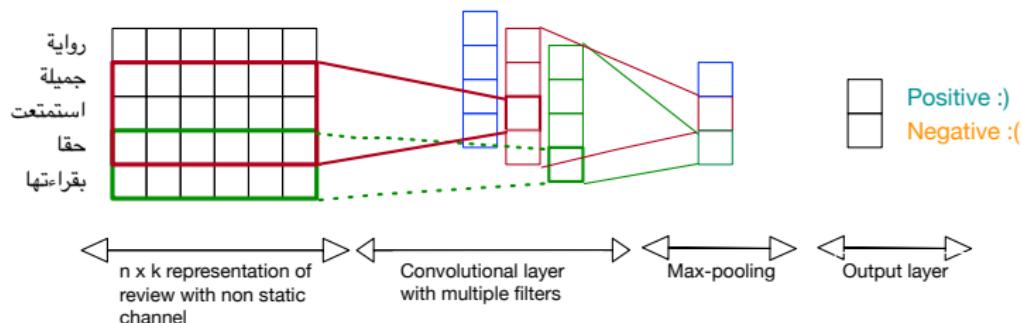
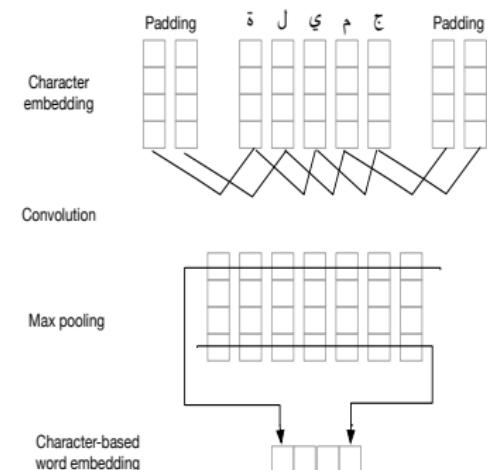
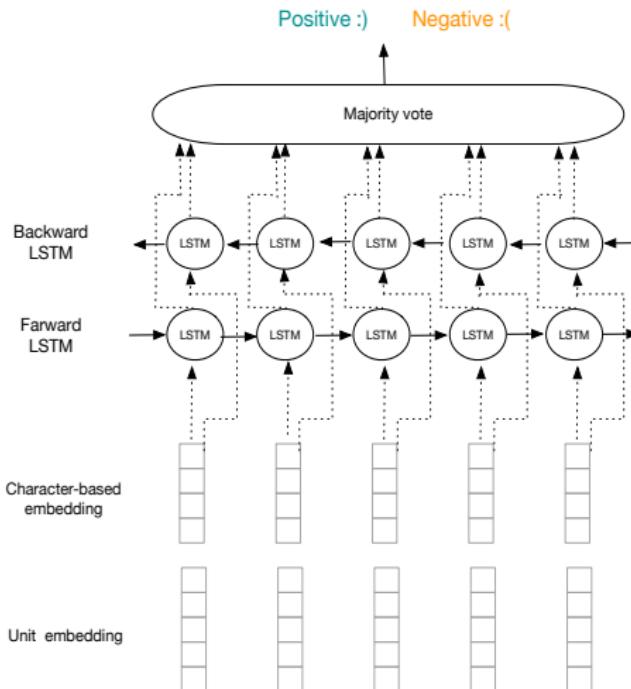


Figure – CNN architecture for a given review.

# BiLSTM-based system

- Architecture similar to [Ma and Hovy, 2016].



## Corpora for embedding training

- *Global* dataset = fusion of existing Arabic SA and newspaper corpora :
  - BRAD [Elnagar et al., 2018b] (510k book reviews).
  - HARD [Elnagar et al., 2018a](373K hotel reviews).
  - LABR [Mahmoud et al., 2014] (only train set composed of 23K book reviews).
  - AbuELkhair [El-Khair, 2016] (5222k news).
- Cleaning and normalisation of *Global* dataset.

| <b>Global</b> | <b>Lexical unit</b> |              |                      |              |                   |             |
|---------------|---------------------|--------------|----------------------|--------------|-------------------|-------------|
|               | <b>Word</b>         | <b>token</b> | <b>token\clitics</b> | <b>lemma</b> | <b>light stem</b> | <b>stem</b> |
| <b>Size</b>   | 3000k               | 1980k        | 1555k                | 950k         | 1997k             | 280K        |

- Embedding dimension = 300 [Dahou et al., 2016, Soliman et al., 2017, Bojanowski et al., 2017]

# Corpus LABR

- LABR corpus *Large-scale Arabic Book Reviews* [Mahmoud et al., 2014] :  
63257 Arabic book reviews (rating scale from 1 to 5) :

| LABR  | *     | **    | ***   | ****   | *****  | total  |
|-------|-------|-------|-------|--------|--------|--------|
| train | 4 195 | 8 554 | 9 561 | 10 136 | 17 960 | 50 606 |
| test  | 1 090 | 2 001 | 2 440 | 3 864  | 3 256  | 12 651 |

- Dataset construction in binary classification framework :
  - Negative reviews : 1 or 2 stars.
  - Positive reviews : 4 or 5 stars.
- Dataset repartition :
  - train set : 90% of considered train LABR
  - validation set : 10% of considered train LABR
  - test set : considered test LABR

# Results

|               |      | CNN  |      |      |      | BiLSTM |      |      |      |
|---------------|------|------|------|------|------|--------|------|------|------|
| Unit          | Emb. | A    | P    | R    | F1   | A      | P    | R    | F1   |
| word          | w2v  | 91.1 | 85.7 | 76.3 | 80.7 | 90.9   | 83.5 | 78.3 | 80.8 |
|               | ft   | 91.2 | 85.2 | 77.2 | 81   | 90.9   | 83.1 | 79   | 81   |
| token         | w2v  | 91.2 | 85.8 | 76.7 | 81   | 90.8   | 83.8 | 77.3 | 80.4 |
|               | ft   | 91.2 | 85.8 | 76.7 | 81   | 90.7   | 83.8 | 76.7 | 80.1 |
| token\clitics | w2v  | 91.2 | 86.4 | 76.0 | 80.9 | 90.8   | 83.8 | 77.2 | 80.4 |
|               | ft   | 91.2 | 85.7 | 76.7 | 80.9 | 90.9   | 84.6 | 76.7 | 80.4 |
| lemma         | w2v  | 91.5 | 85.8 | 78   | 81.7 | 91     | 84.3 | 76.9 | 80.4 |
|               | ft   | 91.4 | 87   | 76.4 | 81.4 | 91     | 83.6 | 78.5 | 81   |
| light stem    | w2v  | 91.2 | 85.6 | 76.8 | 81   | 90.8   | 83   | 78   | 80.4 |
|               | ft   | 91.4 | 86.3 | 76.8 | 81.3 | 90.7   | 83.3 | 77.4 | 80.3 |
| stem          | w2v  | 89   | 81.2 | 69.8 | 75.1 | 89.3   | 80.2 | 73.9 | 76.9 |
|               | ft   | 88.8 | 81.3 | 69   | 74.6 | 89.1   | 79.9 | 73.4 | 76.5 |

**Table –** System evaluations with Arabic-specific embeddings (A = accuracy , P = precision, R = recall, F1 = F1 measure), where 1<sup>st</sup> and 2<sup>nd</sup> best performances.

# Results

|               |      | CNN  |      |      |      | BiLSTM |      |      |      |
|---------------|------|------|------|------|------|--------|------|------|------|
| Unit          | Emb. | A    | P    | R    | F1   | A      | P    | R    | F1   |
| word          | w2v  | 91.1 | 85.7 | 76.3 | 80.7 | 90.9   | 83.5 | 78.3 | 80.9 |
|               | ft   | 91.2 | 85.2 | 77.2 | 81   | 90.9   | 83.1 | 79   | 81   |
| token         | w2v  | 91.2 | 85.8 | 76.7 | 81   | 90.8   | 83.8 | 77.3 | 80.4 |
|               | ft   | 91.2 | 85.8 | 76.7 | 81   | 90.7   | 83.8 | 76.7 | 80.1 |
| token\clitics | w2v  | 91.2 | 86.4 | 76.0 | 80.9 | 90.8   | 83.8 | 77.2 | 80.4 |
|               | ft   | 91.2 | 85.7 | 76.7 | 80.9 | 90.9   | 84.6 | 76.7 | 80.4 |
| lemma         | w2v  | 91.5 | 85.8 | 78   | 81.7 | 91     | 84.3 | 76.9 | 80.4 |
|               | ft   | 91.4 | 87   | 76.4 | 81.4 | 91     | 83.6 | 78.5 | 81   |
| light stem    | w2v  | 91.2 | 85.6 | 76.8 | 81   | 90.8   | 83   | 78   | 80.4 |
|               | ft   | 91.4 | 86.3 | 76.8 | 81.3 | 90.7   | 83.3 | 77.4 | 80.3 |
| stem          | w2v  | 89   | 81.2 | 69.8 | 75.1 | 89.3   | 80.2 | 73.9 | 76.9 |
|               | ft   | 88.8 | 81.3 | 69   | 74.6 | 89.1   | 79.9 | 73.4 | 76.5 |

**Table – System evaluations with Arabic-specific embeddings (A = accuracy , P = precision, R = recall, F1 = F1 measure), where 1<sup>st</sup> and 2<sup>nd</sup> best performances.**

# Results

|               |      | CNN  |      |      |      | BiLSTM |      |      |      |
|---------------|------|------|------|------|------|--------|------|------|------|
| Unit          | Emb. | A    | P    | R    | F1   | A      | P    | R    | F1   |
| word          | w2v  | 91.1 | 85.7 | 76.3 | 80.7 | 90.9   | 83.5 | 78.3 | 80.8 |
|               | ft   | 91.2 | 85.2 | 77.2 | 81   | 90.9   | 83.1 | 79   | 81   |
| token         | w2v  | 91.2 | 85.8 | 76.7 | 81   | 90.8   | 83.8 | 77.3 | 80.4 |
|               | ft   | 91.2 | 85.8 | 76.7 | 81   | 90.7   | 83.8 | 76.7 | 80.1 |
| token\clitics | w2v  | 91.2 | 86.4 | 76.0 | 80.9 | 90.8   | 83.8 | 77.2 | 80.4 |
|               | ft   | 91.2 | 85.7 | 76.7 | 80.9 | 90.9   | 84.6 | 76.7 | 80.4 |
| lemma         | w2v  | 91.5 | 85.8 | 78   | 81.7 | 91     | 84.3 | 76.9 | 80.4 |
|               | ft   | 91.4 | 87   | 76.4 | 81.4 | 91     | 83.6 | 78.5 | 81   |
| light stem    | w2v  | 91.2 | 85.6 | 76.8 | 81   | 90.8   | 83   | 78   | 80.4 |
|               | ft   | 91.4 | 86.3 | 76.8 | 81.3 | 90.7   | 83.3 | 77.4 | 80.3 |
| stem          | w2v  | 89   | 81.2 | 69.8 | 75.1 | 89.3   | 80.2 | 73.9 | 76.9 |
|               | ft   | 88.8 | 81.3 | 69   | 74.6 | 89.1   | 79.9 | 73.4 | 76.5 |

**Table –** System evaluations with Arabic-specific embeddings (A = accuracy , P = precision, R = recall, F1 = F1 measure), where 1<sup>st</sup> and 2<sup>nd</sup> best performances.

# Results

|               |      | CNN  |      |      |      | BiLSTM |      |      |      |
|---------------|------|------|------|------|------|--------|------|------|------|
| Unit          | Emb. | A    | P    | R    | F1   | A      | P    | R    | F1   |
| word          | w2v  | 91.1 | 85.7 | 76.3 | 80.7 | 90.9   | 83.5 | 78.3 | 80.8 |
|               | ft   | 91.2 | 85.2 | 77.2 | 81   | 90.9   | 83.1 | 79   | 81   |
| token         | w2v  | 91.2 | 85.8 | 76.7 | 81   | 90.8   | 83.8 | 77.3 | 80.4 |
|               | ft   | 91.2 | 85.8 | 76.7 | 81   | 90.7   | 83.8 | 76.7 | 80.1 |
| token\clitics | w2v  | 91.2 | 86.4 | 76.0 | 80.9 | 90.8   | 83.8 | 77.2 | 80.4 |
|               | ft   | 91.2 | 85.7 | 76.7 | 80.9 | 90.9   | 84.6 | 76.7 | 80.4 |
| lemma         | w2v  | 91.5 | 85.8 | 78   | 81.7 | 91     | 84.3 | 76.9 | 80.4 |
|               | ft   | 91.4 | 87   | 76.4 | 81.4 | 91     | 83.6 | 78.5 | 81   |
| light stem    | w2v  | 91.2 | 85.6 | 76.8 | 81   | 90.8   | 83   | 78   | 80.4 |
|               | ft   | 91.4 | 86.3 | 76.8 | 81.3 | 90.7   | 83.3 | 77.4 | 80.3 |
| stem          | w2v  | 89   | 81.2 | 69.8 | 75.1 | 89.3   | 80.2 | 73.9 | 76.9 |
|               | ft   | 88.8 | 81.3 | 69   | 74.6 | 89.1   | 79.9 | 73.4 | 76.5 |

**Table – System evaluations with Arabic-specific embeddings (A = accuracy , P = precision, R = recall, F1 = F1 measure), where 1<sup>st</sup> and 2<sup>nd</sup> best performances.**

# Results

|               |      | CNN  |      |      |      | BiLSTM |      |      |      |
|---------------|------|------|------|------|------|--------|------|------|------|
| Unit          | Emb. | A    | P    | R    | F1   | A      | P    | R    | F1   |
| word          | w2v  | 91.1 | 85.7 | 76.3 | 80.7 | 90.9   | 83.5 | 78.3 | 80.8 |
|               | ft   | 91.2 | 85.2 | 77.2 | 81   | 90.9   | 83.1 | 79   | 81   |
| token         | w2v  | 91.2 | 85.8 | 76.7 | 81   | 90.8   | 83.8 | 77.3 | 80.4 |
|               | ft   | 91.2 | 85.8 | 76.7 | 81   | 90.7   | 83.8 | 76.7 | 80.1 |
| token\clitics | w2v  | 91.2 | 86.4 | 76.0 | 80.9 | 90.8   | 83.8 | 77.2 | 80.4 |
|               | ft   | 91.2 | 85.7 | 76.7 | 80.9 | 90.9   | 84.6 | 76.7 | 80.4 |
| lemma         | w2v  | 91.5 | 85.8 | 78   | 81.7 | 91     | 84.3 | 76.9 | 80.4 |
|               | ft   | 91.4 | 87   | 76.4 | 81.4 | 91     | 83.6 | 78.5 | 81   |
| light stem    | w2v  | 91.2 | 85.6 | 76.8 | 81   | 90.8   | 83   | 78   | 80.4 |
|               | ft   | 91.4 | 86.3 | 76.8 | 81.3 | 90.7   | 83.3 | 77.4 | 80.3 |
| stem          | w2v  | 89   | 81.2 | 69.8 | 75.1 | 89.3   | 80.2 | 73.9 | 76.9 |
|               | ft   | 88.8 | 81.3 | 69   | 74.6 | 89.1   | 79.9 | 73.4 | 76.5 |

**Table – System evaluations with Arabic-specific embeddings (A = accuracy , P = precision, R = recall, F1 = F1 measure), where 1<sup>st</sup> and 2<sup>nd</sup> best performances.**

# System combination

Combination protocols :

Oracle (ideal combination) and Consensus (agreement).

| Unit embeddings | System prediction     | Ref | Oracle | Consensus |
|-----------------|-----------------------|-----|--------|-----------|
| All \ stem      | ● ● ● ● ● ○ ● ● ● ● ● | ○   | ○ ✓    | reject    |
|                 | ● ● ● ● ● ● ● ● ● ●   | ○   | ○ ✗    | ○ ✗       |
|                 | ○ ○ ○ ○ ○ ○ ○ ○ ○ ○   | ○   | ○ ✓    | ○ ✓       |
| 2 best          | ○ ○                   | ○   | ○ ✓    | reject    |
|                 | ○ ○                   | ○   | ○ ✓    | ○ ✓       |
|                 | ● ●                   | ○   | ○ ✗    | ○ ✗       |

# System combination

Combination protocols :

Oracle (ideal combination) and Consensus (agreement).

| Unit embeddings | System prediction       | Ref | Oracle | Consensus |
|-----------------|-------------------------|-----|--------|-----------|
| All \ stem      | ● ● ● ● ● ● ● ● ● ● ● ● | ○   | ○ ✓    | reject    |
|                 | ● ● ● ● ● ● ● ● ● ● ●   | ○   | ○ ✗    | ○ ✗       |
|                 | ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○   | ○   | ○ ✓    | ○ ✓       |
| 2 best          | ○ ○                     | ○   | ○ ✓    | reject    |
|                 | ○ ○                     | ○   | ○ ✓    | ○ ✓       |
|                 | ● ●                     | ○   | ○ ✗    | ○ ✗       |

|            |           | CNN      |          | BiLSTM   |          |
|------------|-----------|----------|----------|----------|----------|
|            | Protocol  | Coverage | Accuracy | Coverage | Accuracy |
| All \ stem | Oracle    | 100%     | 100%     | 100%     | 100%     |
|            | Consensus | 86.57%   | 100%     | 81.70%   | 100%     |
| 2 Best     | Oracle    | 100%     | 92.3%    | 100%     | 92.5%    |
|            | Consensus | 98.25%   | 92.2%    | 96.96%   | 92.2%    |

## 2 Best CNN Combination frame : non-consensus analysis

- Number of non-consensus reviews = 146 reviews :
  - 68 positive (19 with 5\* and 49 with 4\*)
  - 78 negative (20 with 1\* et 58 with 2\*).

|                   | Examples   |
|-------------------|--|
| Mixed review      | المواضيع تقليدية لكن الأسلوب جميل                    |
|                   | اسلوب قوي واحذاث ضعيفة                               |
| Author evaluation | على علي زيدان ان يترك الكتابة ليقى كيبيا بعيون الناس |
|                   | لا تعليق، فاي تعليق يقلل من قيمة الكاتب              |
| Star number       | لقد كنت كريمة جدا باعطاء نجمتين                      |
|                   | انقصت نجمة في التقييم                                |

## Conclusion

- Arabic sentiment analysis framework.
- Specificity of Arabic language : agglutination and morphological richness.
- 12 Arabic-specific embedding sets :  $unit\_Emb$ , where
  - $unit \in \{\text{word, token, token}\backslash\text{clitics, lemma, light stem, stem}\}$
  - $Emb \in \{\text{w2v, ft}\}$ .
- 2 neural systems : CNN-based and BiLSTM-based.

## Conclusion

- Lemma  $\approx$  the most appropriate lexical unit for Arabic SA.
- Best performance in this work = 91.5% of accuracy
- Accuracy of [Barhoumi et al., 2018] = 89.3%.
  - Choice of document length and padding/truncating type = + 1
  - Similar domain corpora for training embeddings = + 0.8
  - Lemma embeddings = + 0.4
- Note that :
  - Size (word embedding space) = 3000k
  - Size (lemma embedding space) = 950k

## Future works

- Evaluation of different Arabic-specific embeddings in other NLP tasks (POS tagging, NER, syntactic and semantic analogies).
- Sentiment embeddings [Yu et al., 2017].
- Contextualized embeddings ELMO [Peters et al., 2018].

Thank you for your  
attention :)

# References

-  Al-Smadi, M., Talafha, B., Al-Ayyoub, M., and Jararweh, Y. (2018).  
Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews.  
*International Journal of Machine Learning and Cybernetics*, pages 1–13.
-  Alayba, A. M., Palade, V., England, M., and Iqbal, R. (2018).  
Improving sentiment analysis in arabic using word representation.  
In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 13–18. IEEE.
-  Barhoumi, A., Camelion, N., and Estève, Y. (2018).  
Des représentations continues de mots pour l'analyse d'opinions en arabe : une étude qualitative.  
In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018)*, Rennes, France.
-  Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016).  
Enriching word vectors with subword information.  
*arXiv preprint arXiv :1607.04606*.
-  Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017).  
Enriching word vectors with subword information.  
*Transactions of the Association for Computational Linguistics*, 5 :135–146.
-  Dahou, A., Elaziz, M. A., Zhou, J., and Xiong, S. (2019).  
Arabic sentiment classification using convolutional neural network and differential evolution algorithm.  
*Computational intelligence and neuroscience*, 2019.
-  Duan, A., Xiong, S., Zhou, J., Hu, H., He, J., Ma, H., and Dong, P. (2016).