



HAL
open science

Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?

Lisa Medrouk, Anna Pappa

► **To cite this version:**

Lisa Medrouk, Anna Pappa. Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?. 2018 International Joint Conference on Neural Networks (IJCNN), Jul 2018, Rio de Janeiro, Brazil. pp.1-6, 10.1109/IJCNN.2018.8489613 . hal-03623944

HAL Id: hal-03623944

<https://hal.science/hal-03623944>

Submitted on 29 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DRAFT

¹To cite this article : L. Medrouk and A. Pappa, "Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?," 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-6, doi: 10.1109/IJCNN.2018.8489613.

Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?

Lisa Medrouk

LIASD

University Paris 8

France

lmedrouk@ai.univ-paris8.fr

Anna Pappa

LIASD

University Paris 8

France

ap@ai.univ-paris8.fr

Abstract—In this article we introduce an empirical study of multilingual and multi-topic opinion classification. The particularity relies on the reviews that are written in different languages and refer to different but semantically close topics: *Restaurants and Hotels*. Our key objective is to emphasize the ability of a deep learning model to establish the sentiment polarity of reviews and topics Classification in a multilingual environment without any *prior* knowledge. For this work, we use unstructured text data, collected from the web, written in French, English and Greek (a less opinion-present language). We incorporate corpus-based input is *raw*, used without any pre-processing, translation, annotation nor additional knowledge features. For the machine learning approach, we use two different deep neural networks, Convolutional Neural Networks (CONVNETS) and Recurrent Neural Networks (RNNs). The learning model exploits n-gram level information, and achieves high accuracy for sentiment polarity and topics classification according to the experimental tests and results. From our hypothesis, we argue that the multilingual environment composed of reviews in semantically close domains, does not impact the network performance, and lead us to deduce that semantic features extraction with ConvNets and RNNs are language and context independent. Following these results, we tend to promote the inception of simple yet powerful approach for feeding deep networks in multilingual context.

Index Terms—Deep Learning, Multi-Domain Classification, Multilingual Classification, Opinion Analysis, ConvNets, LSTM, Raw Text Corpus.

I. INTRODUCTION

THE challenge with multilingual opinion analysis is always relevant. The Web2.0 offers a wide range of reviews plate-forms, from personal blogs, to online products review. The Web reviews data can be compared to the Tower of Babel, in sense of their abundance, and the fact that they are written in different languages and covering multiple domains. Despite this complexity the user still wants a simple way to extract, understand and analyze 'what is said' regardless the language used.

Deep Neural Networks (DNN) have gained attention and popularity over the last years, and have been widespread and yielding state-of-the-art results on various Natural Language Processing (NLP) tasks such as Part-Of-Speech (POS) tagging [1], morphological tagging and disambiguation [2], language

modeling [3]. DNN has also establish remarkable results in challenging NLP classification tasks such as Sentiment Analysis.

Working on a multilingual environment implies facing the lack of resources such as labeled data in "poor" languages, building these resources for all the targeted languages is time consuming and expensive. In order to address this issue, we enhance the simple yet effective approach introduced in [4] using neural language models and deep learning techniques in a multilingual environment, without relying on any prior knowledge, such as lexicon, bilingual dictionary [5], language switch indication nor using a cross-language.

In addition to the multilingual challenge we add an extra complexity with the semantically *close domain inputs*. Restaurants and hotels reviews are in some cases very similar, as hotels offers food facilities, and in both cases users can review similar services which may lead to some confusion for classification tasks, even for humans.

Our purpose is to answer this question: Does Deep networks really need the extra pre-training hassle to process multilingual datas?

The remainder of this paper is structured as follows. In section II, we start by presenting the related literature, in section III and section IV we describe our method with the proposed models, section V presents the corpus and the conducted experimentation, and finally, we discuss the results in section VI and we draw our conclusions in section VII by answering the question painted in our title.

II. RELATED WORK

Prior studies have successfully used deep networks for mono-lingual task classification in general, and more specifically, for sentiment and opinion analysis [6], showing that deep networks have powerful feature learning abilities. A quick look at the main language used unsurprisingly reveals that English is still the most used language in reviews, as it remains the most widely used Lingua Franca in international communications [7]. Nevertheless, the user participation barometer of web 2.0 technologies raises the other languages presence in

significantly high scores especially in social media (for Twitter for example, the most used second language is Portuguese and the third one is Japanese¹).

For the other languages, [8] compared ConvNets and LSTM models for sentiment analysis of Russian tweets, [9] explored four different architectures : Deep Belief Networks (DBN) and a combined Auto Encoder with DBN for text sentiment analysis in Arabic, and [10] provided a solution to multilingual sentiment of poor languages classification using deep learning through English translation.

The multilingual modeling approaches use hybrid models, often combining trained HMM models to DNN architecture, but access to dictionaries and language models is necessary [11]. Hybrid approaches for learning embeddings use jointly incorporating corpus-based evidence and lexical or semantic resources (structured knowledge). This cooperation between structured knowledge and corpus-based approaches has proven rich for knowledge embeddings, useful for sentiment analysis, but still requires other knowledge sources [12].

While most text classification are mono lingual centring, we believe that a multilingual model is more accurate as it reflects current online reviews nature trends. In the restaurant domain for example, with Europe as location, it's common to find reviews written in English, French, Spanish, Italian, Dutch, or any other language depending mostly on the customer native or more convenient language. We propose a DNN classification model for mixed multilingual and multi-topic corpus, without any lexicon or other structured knowledge.

III. OUR METHOD

We attempt in this study to highlight deep network abilities in a multilingual and mutli-topic context. Our corpus is composed of text data : reviews and opinions, automatically collected from the web, targeting hotels and restaurants regardless of the location and the language.

The reviews written in different languages are used without any prior knowledge of language switching nor transition. In our case, we expect association of meanings to sentiments in a semantically close context. Given this premise, we decide to have a unique input composed of a balanced shuffled number of reviews written in three heterogeneous languages: An Anglo-Saxon, a Roman language and a Greek language.

We report a series of experiments using *ConvNets* [13] for their ability to extract local features neighbourhood and a specific type of RNNs, Long short-term memory networks *LSTMs* [14] for their capacity to learn long term dependencies from sequences.

We focus on two common NLP tasks: Domain Classification and Sentiment Polarity Classification. We choose to work using a novel feeding approach that consist of mixing indifferently poor and rich languages and topics in one single stream. The raw mix input is presented to a deep neural net model that already showed great NLP tasks potential using single languages and mono topics. With one binary label

and providing the network models with the most pertinent hyper-parameters, we attempted to study the models features extraction capacities using this raw complex input without extra modules nor translation for both classification tasks.

IV. THE PROPOSED MODELS

This study investigates two different types of deep networks : RNNs and ConvNets. Using a deep network allow us to tackle the multilingual problem without focusing on special features and therefore, have the opportunity to let the model learn it's appropriate features.

Our first model uses ConvNets and the second one uses LSTMs. Both models are followed by non-linear layers, in order to allow the network to learn non-linear decision boundaries. The models are fed with a unique input composed of a shuffled number of reviews in English, French and Greek covering restaurants and hotels domains.

A. Convolutional Neural Networks

ConvNets have originally been applied for image processing and are the core of most Computer Vision systems today, they have since, proved their efficiency in NLP and are now widely used for text classification. ConvNets have successfully been used by [15] for a variety of sentence classification tasks, by [16] who proposed a new deep convolutional neural network that exploits character to sentence level information to perform sentiment analysis of short texts, by [17] who used a dynamic convolutional neural network (DCNN) for semantic modeling of sentences, and by [18] who used a ConvNets for an aspect based sentiment analysis task and also by [19] who used a recurrent convolutional neural networks for text classification. Just like traditional neural networks, a ConvNets consist of neurons that have learnable weights, the difference comes from the learning method, as for convNets this operation is named convolution. A ConvNets is usually composed of one or more than one convolution layers on which is applied a nonlinear activation function. The aim of the convolutional layer is to extract discriminative patterns, and a convolution operation involves a convolutional kernel i.e. a filter with a specific window size sliding over a window of words represented by a vector in order to extract different features. This operation is applied to every n-grams of the input text, producing a feature map, trying to extract the most important n-gram.

For a filter represented by a weight vector $W \in R^{h \times d}$; a convolution on a sentence of n consecutive words expressed by the sentence matrix $X \in R^{s \times d}$ can be expressed as follows:

$$c_i = f(W \cdot X_{i:i+n-1} + b_f) \quad (1)$$

where (\cdot) is the Hadamart product between the sub-matrix sentence from i to j and the filter, b being the bias. f a non linear function and c the feature map for the filter $c \in R^{s-n+1}$. Our ConvNet model is composed by a single convolutional layer, followed by the activation function *ReLU* (*Rectified Linear Unit*) which simply threshold the output at 0 when the input is negative and is defined by $F(x) = \max(x, 0)$.

The convolutional layer is interleaved with a pooling layer

¹<https://semioast.com/en/publications>

which shrinks neighbouring features. We apply a MaxPooling that extracts the maximum value for each filter by simply returning the *max* value from each previous maps captured features with the highest value, which is simply the most informative n-gram for the following steps.

B. Long Short Term Memory

Unlike feedforward neural networks, RNNs standing for *Recurrent neural networks* can use their internal memory to handle a sequence of inputs, which make them by nature the dominant approach for many NLP tasks. In RNNs, predictions are made sequentially; the hidden layer from one prediction feeds the hidden layer of the following prediction, implying a sort of memory across layers.

[20] applied a RNN architecture for a sentiment prediction task at phrase and sentence level. In practice, RNNs have one drawback as they can't really handle long sequences, as they start to "forget" after few sequences due to two causes, the vanishing and the exploding gradients descent. To address this issue that commonly appear during the RNNs training phase, [21] proposed an extended model named LSTM standing for *Long Short Term Memory*. The LSTM architecture is an augmented version of the RNN with additional gates introduced in order to monitor the data flow : a *forget gate* to control whether to forget the current state; an *input gate* to indicate if it should read the input; an *output gate* to control whether to output the state. These small enhanced operations are very efficient at learning problems with long range temporal dependencies. The LSTM model has already successfully been used for detecting sentiment polarity in [22].

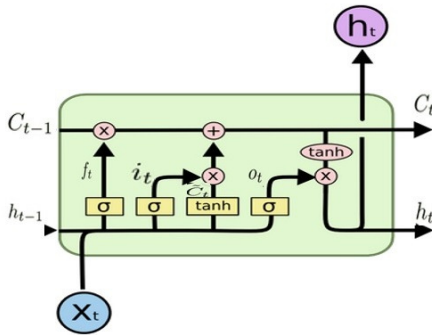


Fig. 1. LSTM Block, image sourced from [23]

The illustration of an unrolled LSTM block in Figure 1 represent h_{t-1} the output for the preceding block, C_{t-1} stands for the memory from the precedent block, X_t is the entry vector. h_t is output of the current block, C_t is the memory of the actual block.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$c_t = f_t \odot C_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

$$o_t = \sigma(W_{io} \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \odot \tanh(C_t) \quad (7)$$

$W_j \in R^{n \times m}$ is a weight matrix, $b_j \in R^n$ bias vector for $j \in \{i, f, c, o\}$. The LSTM has four gates, an input gate i_t who decide which value to update, a forget gate f_t that outputs a number between 0 and 1, where 1 means *keep* and 0 means *forget* in order to keep only relevant information from the memory. An output gate o_t , all calculated by a sigmoid function over the input x_t and the preceding hidden state h_{t-1} . \tilde{c}_t is a temporary state that creates a vector of new candidate values. \odot in the equation stands for the element-wise multiplication, c_t is the new cell state, o_t, h_t stands for which part of the cell state to output, and finally \tilde{c}_t, i_t the new information to remember and to output to the next cell.

Basically, these gates decide what to keep in and what to delete from the memory, then, they combine the new state with the previous memory and the input. Our LSTM model uses a LSTM layer of size 70, followed by a sigmoid classification layer, defined by the activation function:

$$f(x) = (1 + e^{-x})^{-1} f : R \rightarrow [0, 1]. \quad (8)$$

Our LSTM model is composed by a single LSTM layer, followed by a fully connected layer before prediction.

V. CORPUS AND EXPERIMENTATION

For testing the proposed experiments, we constitute different combination of corpora composed of French, English and Greek reviews covering restaurants and hotels domains with positive and negative polarities, all automatically extracted in their original language to avoid any kind of noise due to a translation system. The main corpus is composed of 92K (K being 1000 reviews). In order to compare the mono and multilingual approaches, we conducted our tests on both the main multilingual corpus and on it's languages subsets :

- 1) Only French
- 2) Only English
- 3) French and English
- 4) English, French and Greek

No pre-trained word embedding [24] [25] or any other text treatment has been performed on the corpus. The input sentences are simply lower cased, tokenized and converted into a matrix $X \in R^{s \times d}$, where the rows are d-dimensional word vectors for each token of size 34 for our models, and s denotes the sentence length, set to 500 in our case which is a bit higher than the average size of the collected reviews.

Finding the optimal hyperparameters with the right choice of network architecture, and effective regularization is complex and still based on experience. To find the effective

hyperparameters of both (ConvNet and LSTM) of our experimental models, we conducted a pre-trained experimental phase, selecting our optimal parameters by testing a range of hyperparameter, testing one from the list at a time while keeping the other ones unchanged. The experimental tests has been applied the French-English multi-topic input. Every test has been run for 4 epochs, examining the losses of our tests sets and picked the hyperparameters of the most balance results. Table I highlights our optimal choices that has been applied to all our experiments. The models are trained through *stochastic gradient descent* over mini-batches pack of length 64, via *backpropagation* to minimize the *binary cross entropy* loss using Adam optimizer [26].

For regularization we mainly employ dropouts [27], that randomly omit part of the feature detectors during training phases which prevent co-adaptation.

TABLE I
OPTIMAL HYPERPARAMETERS AND REGULARIZATION PICKED TO TRAIN THE EXPERIMENTAL MODELS

Hyperparameters	Experiment Range	Choice
Word vectors dimension	25-120	34
ConvNet number of filters	30-100	64
ConvNet kernel size	3-7	5
MaxOverPooling	2-4	4
ConvNet dropout rate	10-45	40
LSTM dimensionality	50-120	70
LSTM dropout rate	10-45	20
Optimization	Adam, RMSprop	Adam

We used in both models a single level architecture, for the ConvNets model we choose a single convolution layer, as it has been proved by [15] that a single layer model perform equally well in comparison with a multi-layer model. We conducted the same experiments with LSTM models and observed the same result.

VI. RESULTS AND DISCUSSION

We conducted our experiments on 4 datasets for each classification task. the aim of the benchmark is to report enough results to compare multilingual against monolingual performance.

Table II, presents the experiment the of a ConvNets model applied to a topic classification task using the main multilingual corpus and its subsets. Table III shows the same experiment task using the LSTM model. Both experiments have been performed using the following inputs decomposition:

- 1) A dataset of 91,816 English, French and Greek Restaurants and Hotels reviews trained on 64,272 multilingual samples and tested on 27,544 multilingual samples.
- 2) A dataset of 83,980 English and French Restaurants and Hotels reviews trained on 58,786 bilingual samples and tested on 25,194 bilingual samples.
- 3) A dataset of 57,176 English Restaurants and Hotels reviews trained on 40024 English samples and validated on 17,152 English samples.

- 4) A dataset of 26,804 French Restaurants and Hotels reviews trained on 18,763 French samples and validated on 8,041 French samples.

TABLE II
ACCURACY RESULTS OF THE CONVNETS MODEL FOR TOPIC CLASSIFICATION TASK OF THE MIX OF HOTELS AND RESTAURANTS REVIEWS IN FRENCH, ENGLISH AND GREEK HIGHLIGHTING THE FACT THAT THE MIX LANGUAGES INPUT PERFORMED EQUALLY WELL IN COMPARISON WITH A MONOLINGUAL INPUT

Languages	Accuracy
English and French and Greek	98.42
English and French	98.42
French	98.94
English	98.16

TABLE III
ACCURACY RESULTS OF THE LSTM MODEL FOR TOPIC CLASSIFICATION TASK OF A MIX OF HOTELS AND RESTAURANTS REVIEWS IN FRENCH, ENGLISH AND GREEK HIGHLIGHTING EQUAL RESULTS FOR THE MIX LANGUAGES INPUT IN COMPARISON WITH A MONOLINGUAL INPUT

Languages	Accuracy
English and French and Greek	98.43
English and French	98.37
French	99.05
English	97.81

The results accuracy of the domain classification task of each model are all very high, exceeding or equal to 98% for both models, and more importantly highlighting the fact that the mix languages input performed equally well in comparison with a monolingual using both models. The results also reflect the fact that both ConvNets and LSTMs models performed equally well using single bi and multilingual inputs.

Table IV, presents the experiment of the ConvNets model applied to the opinion classification task using the different datasets composed of mono and multilingual reviews of hotels and restaurants. Table V shows the same experiment task using the LSTM model.

TABLE IV
ACCURACY RESULTS OF THE CONVNETS MODEL FOR OPINION CLASSIFICATION TASK OF A MIX OF HOTELS AND RESTAURANTS REVIEWS IN FRENCH, ENGLISH AND GREEK, HIGHLIGHTING EQUAL RESULTS FOR THE MIX LANGUAGES INPUT AND THE MONOLINGUAL INPUT

Languages	Accuracy
English and French and Greek	91.25
English and French	91.74
French	92.76
English	91.27

The opinion classification task accuracy of each model is above or equal to 91%, and more importantly, it shows that the multilingual input performed equally well in comparison to bi and monolingual input using both models.

The models performed better for domain classification 98% tasks in comparison to the polarity prediction 91% for all the languages with the different composition due probably to languages inherent nature and highlighting the fact the opinions are more complex to classify than the topics.

TABLE V
ACCURACY RESULTS OF THE LSTM MODEL FOR OPINION CLASSIFICATION TASK OF A MIX OF HOTELS AND RESTAURANTS REVIEWS IN FRENCH, ENGLISH AND GREEK, HIGHLIGHTING EQUAL RESULTS FOR THE MIX LANGUAGES INPUT AND THE MONOLINGUAL INPUT

Languages	Accuracy
English and French and Greek	91.27
English and French	91.16
French	92.68
English	90.94

Digging further, we tried to identify some *false positives* and *false negatives* reviews from the multilingual corpus experiments. The false positives rate is the proportion of all the negative reviews that still yield positive test outcomes and complementary, the false negatives rate is the proportion of the positive reviews which yield negative test outcomes with the test set. Table VI and VII highlights the classification reports for both tasks with multilingual corpus. Table VI highlights the results on 27,544 validation sets, from which we obtained 217 false positives reviews, and 218 false negatives reviews.

TABLE VI
CLASSIFICATION REPORT OF THE CNN MODEL USING MULTILINGUAL REVIEWS FOR THE TOPICS CLASSIFICATION TASK

Topic	Number of data	Precision	Recall	F_1 score
Restaurant	18653	0.99	0.99	0.99
Hotel	8891	0.98	0.98	0.98

TABLE VII
CLASSIFICATION REPORT OF THE CNN MODEL USING MULTILINGUAL REVIEWS FOR THE OPINION CLASSIFICATION TASK

Label	Number of data	Precision	Recall	F_1 score
Negative	12688	0.91	0.90	0.90
Positive	14856	0.92	0.92	0.92

Table VIII highlights the classification report for the opinion classification task on 26,804 French reviews with results on 8,041 validation sets, from which we obtained 323 false positives reviews, and 259 false negatives reviews.

TABLE VIII
CLASSIFICATION REPORT OF THE CNN MODEL USING FRENCH REVIEWS FOR THE OPINION CLASSIFICATION TASK

Label	Number of data	Precision	Recall	F_1 score
Negative	3746	0.92	0.93	0.92
Positive	4295	0.94	0.92	0.93

Table IX highlights the classification report for the topic classification task on 57,176 English reviews with results on 17,152 validation sets, from which we obtained 161 false positives reviews, and 154 false negatives reviews.

We present below some examples of some misqualified reviews obtained with the ConvNets model using the multilingual corpus. The reviews are delivered in the raw format, as originally retrieved and including the original misspelling or common online reviews errors.

TABLE IX
CLASSIFICATION REPORT OF THE CNN MODEL USING ENGLISH REVIEWS FOR THE TOPIC CLASSIFICATION TASK

Topic	Number of data	Precision	Recall	F_1 score
Restaurant	10634	0.98	0.99	0.99
Hotel	6518	0.98	0.98	0.98

Examples of Domain classification false positive or false negative:

- Restaurant review identified as Hotel review : "after a cruise we headed to maxine's for breakfast as we were staying at the sls hotel." , the word hotel was probably the feature inducing the error. Same remark with word "room" for this review : "went fir a party there. the coral room is amazing and the food and service first class. had a great evening."
- Restaurant review identified as Hotel review : " , ! ! ". This review is a reference to a Greek song and wouldn't have been difficult to classify for non native.
- Hotel review identified as restaurant review : "Agrablement surpris par la qualit du service gnral". Very generic sentence, that can be confusing even for a human reviewer. The same case for this review "we had a good time, but the staff didn't act like they cared enough."
- Hotel review identified as restaurant: "this small hotel, turned out to be an excellent surprise: ideally located close to the opera garnier, guarantees direct access to all city by foot or by metro/bus, is surrounded by innumerable bars, restaurants and theatres, with a good atmosphere in the street." The sentence include both restaurant and hotel which might cause the classification error. Noticing that a human wouldn't misqualified this review.

Examples of sentiment analysis false positives or false negatives reviews for bi-language input:

- Negative review identified as positive: "the first time we were here a year ago we really loved this place. now, one year later, almost everything about our experience here was a big disappointment."
- Negative review identified as positive: "la politesse est un service agrable ce n'est pas las bas." This is an ironic French review.
- Positive review identified as negative: "i love this place for its hidden location. it is very cozy inside and the most important part you never had this delicious pizza before. the only disappointing thing is before they used to give complimentary croissant with nutella. now unfortunately they don't serve them anymore".

We haven't noticed a particular pattern error, while reviewing the different misqualified reviews from all our results. Mono and multilingual corpus seems to address, indifferently, the same kind of errors.

Some of the presented examples highlights the fact that the misqualified reviews correlate with the common issues faced in both NLP classification tasks, like irony detection, or very general context or common words issues.

From all the presented results we can conclude that deep networks can equally perform extremely well with mono and multilingual corpus without extra modules nor extra handcrafted knowledge, which let us lead toward a *No* to answer our title question: deep Networks do not need complex modules for multilingual Sentiment polarity detection and domain classification.

VII. CONCLUSION

In this paper, we introduce a raw multilingual model for domain classification and polarity detection. Contrary to most existing methods, both of our tasks prediction used a unique multilingual and multi-domain dataset source for automatic learning without extra languages modules, nor languages switch or translation. We demonstrate that deep neural networks process multilingual text input and single language text input in the same way without relying on any linguistic knowledge, even when adding extra complexity like very close topics. To highlight our proposal, the input has been tested with two deep neural networks, LSTM and ConvNets. All the empirical results for single and multilingual and multi-topics models for both tasks, showed a quasi similar and remarkable performance, allowing us to conclude that both models extract features indifferently while treating single or multilingual datasets, and therefore implying that we can avoid extra complex *à priori* modules when performing multilingual classification tasks with deep networks.

Future work should explore more poor language effects and use bigger datasets.

REFERENCES

- [1] A. Popov, "Deep learning architecture for part-of-speech tagging with word and suffix embeddings," in *Artificial Intelligence: Methodology, Systems, and Applications*, C. Dichev and G. Agre, Eds. Cham: Springer International Publishing, 2016, pp. 68–77.
- [2] E. Yildiz, C. Tirkaz, H. B. Sahin, M. T. Eren, and O. Sonmez, "A morphology-aware network for morphological disambiguation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2863–2869.
- [3] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, ser. WLM '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 20–28. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2390940.2390943>
- [4] L. Medrouk and A. Pappa, "Deep learning model for sentiment analysis in multi-lingual corpus," in *Neural Information Processing - 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I*, 2017, pp. 205–212. [Online]. Available: https://doi.org/10.1007/978-3-319-70087-8_22
- [5] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 2007, pp. 976–983. [Online]. Available: <http://www.aclweb.org/anthology/P07-1123>
- [6] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 720–728.
- [7] B. Bjrkman, "The pragmatics of english as a lingua franca in the international university: Introduction," *Journal of Pragmatics*, vol. 43, no. 4, pp. 923 – 925, 2011, english as a Lingua Franca. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378216610002717>
- [8] A. K., K. I., T. J., G. A., and T. D. S. K., "Comparison of neural network architectures for sentiment analysis of russian tweets," *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2016*, 2016.
- [9] A. A. Sallab, H. M. Hajj, G. Badaro, R. Baly, W. El-Hajj, and K. B. Shaban, "Deep learning models for sentiment analysis in arabic," in *ANLP@ACL*, 2015.
- [10] P. Singhal and P. Bhattacharyya, "Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification," in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 2016, pp. 3053–3062.
- [11] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*. IEEE, 2013, pp. 7319–7323. [Online]. Available: <http://dblp.uni-trier.de/db/conf/icassp/icassp2013.html#GhoshalSR13>
- [12] H. Nguyen and M.-L. Nguyen, "A deep neural architecture for sentence-level sentiment classification in twitter social networking," *CoRR*, vol. abs/1706.08032, 2017.
- [13] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258. [Online]. Available: <http://dl.acm.org/citation.cfm?id=303568.303704>
- [14] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.
- [15] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.
- [16] C. N. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *COLING*, 2014, pp. 69–78.
- [17] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *CoRR*, 2014.
- [18] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *ACL*, 2015.
- [19] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 2267–2273. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2886521.2886636>
- [20] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013, pp. 1631–1642.
- [21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [22] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *ACL*, 2015.
- [23] C. Olah, "Understanding lstm networks," 2015. [Online]. Available: <http://colah.github.io/posts/2015-08-Understanding-LSTMs>
- [24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: <http://www.aclweb.org/anthology/D14-1162>
- [25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [27] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012.