

TMLab SRPOL at SemEval-2019

Task 8: Fact Checking in Community Question Answering Forums

Piotr Niewiński
Aleksander Wawer
Maria Pszona
Maria Janicka

Fake News Recognition Project @ Samsung R&D Institute Poland



2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics



SemEval-2019: International Workshop on **Semantic Evaluation** Sponsored by SIGLEX and Microsoft

Organizers:

- Jonathan May, ISI, University of Southern California
- Ekaterina Shutova, University of Amsterdam
- Aurelie Herbelot, University of Trento
- Xiaodan Zhu, Queen's University
- Marianna Apidianaki, LIMSI, CNRS, Université Paris-Saclay & University of Pennsylvania
- Saif M. Mohammad, National Research Council Canada

SemEval-2019 Workshop will be held June 6-7, 2019
in Minneapolis, USA, collocated with NAACL-HLT 2019



12 tasks in 5 categories:

- Frame semantics and semantic parsing
- Opinion, emotion and abusive language detection
- Fact vs fiction (Task 8 here)
- Information extraction and question answering
- NLP for scientific applications

SemEval 2019 **Task 8** on Fact-Checking in Community Forums:

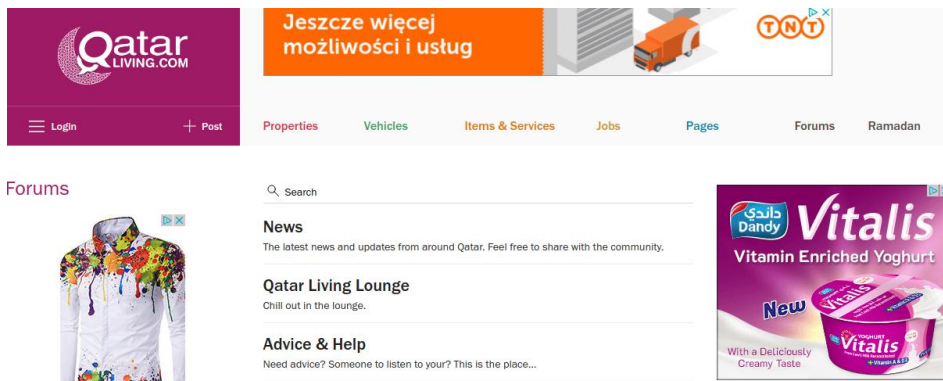
- **Subtask A:** decide whether a question asks for a **factual** information, an **opinion**/advice or is just **socializing**.
- **Subtask B:** decide whether an answer to a factual question is true, false, or does not constitute a proper answer.

SemEval Task 8, Subtask A

Question classification problem, goal is to classify questions from the forum in three categories:

- **Factual:** The question is asking for factual information, which can be answered by checking various information sources, and it is not ambiguous (e.g., "What is Ooredoo customer service number?").
- **Opinion:** The question asks for an opinion or an advice, not for a fact. (e.g., "Can anyone recommend a good Vet in Doha?")
- **Socializing:** Not a real question, but intended for socializing or for chatting. This can also mean expressing an opinion or sharing some information, without really asking anything of general interest. (e.g., "What was your first car?")

Samples come from “Qatar Living” forum: <https://www.qatarliving.com/forum>



The screenshot displays the Qatar Living forum website. At the top left is the logo for Qatar Living.com. Below it are 'Login' and 'Post' buttons. A navigation menu includes 'Properties', 'Vehicles', 'Items & Services', 'Jobs', 'Pages', 'Forums', and 'Ramadan'. A search bar is located below the navigation. The main content area features several sections: 'News' with a sub-header 'The latest news and updates from around Qatar. Feel free to share with the community.', 'Qatar Living Lounge' with the text 'Chill out in the lounge.', and 'Advice & Help' with the text 'Need advice? Someone to listen to your? This is the place...'. On the left side, there is a 'Forums' section with a thumbnail image of a white shirt decorated with colorful paint splatters. On the right side, there is an advertisement for Vitalis Vitamin Enriched Yoghurt, featuring a tub of the product and the text 'New' and 'With a Deliciously Creamy Taste'.

Quiz - examples from data (human performance)

1. why its very hard to get an approval for female working visa here in Qatar?
1. Hi;Can any one tell me a place where i can have a good massage drom philipinies????? yesterday i had a massage in Bio-Bil they charged me 300qr for 01 hour bt it is totally waste... pls advice me if theres any philipinos....
1. Hi; Let's speak today about some transportation devices Words: Baas = Bus SayYara = Car Qetar = Train Safeena = Ark Zawraq= Canoe Quiz 1- How can we say "What is the color of Taxi cars in Doha?" in Arabic? – plz use English letters – 2- What is the English Translation of "Aen Maoqef al-Baas?" Have a nice day :)
1. Red Light Jumped,Red Light Jumped. Ooopps !!!! I just crossed the Stop line at Crazy signal; But did not crossed the intersection . The camera Lights flashed 2 times and I stood there only. Will it be fined adn how much????????
1. Qatar CV format? Hey everyone; I'm applying for a job in Qatar; and I'm wondering how to format my CV. I already have a suitable CV for the UK; so please could someone let me know if there's any extra information that I need to add to my CV; to meet the standards in Qatar? Thanks in advance! :)

Quiz - examples from data (human performance)

1. why its very hard to get an approval for female working visa here in Qatar?

Factual

1. Hi;Can any one tell me a place where i can have a good massage drom philipinies????? yesterday i had a massage in Bio-Bil they charged me 300qr for 01 hour bt it is totally waste... pls advice me if theres any philipinos....

Opinion

1. Hi; Let's speak today about some transportation devices Words: Baas = Bus SayYara = Car Qetar = Train Safeena = Ark Zawraq= Canoe Quiz 1- How can we say "What is the color of Taxi cars in Doha?" in Arabic? – plz use English letters – 2- What is the English Translation of "Aen Maoqef al-Baas?" Have a nice day :)

Socializing

1. Red Light Jumped,Red Light Jumped. Ooopps !!!! I just crossed the Stop line at Crazy signal; But did not crossed the intersection . The camera Lights flashed 2 times and I stood there only. Will it be fined adn how much????????

Factual

1. Qatar CV format? Hey everyone; I'm applying for a job in Qatar; and I'm wondering how to format my CV. I already have a suitable CV for the UK; so please could someone let me know if there's any extra information that I need to add to my CV; to meet the standards in Qatar? Thanks in advance! :)

Opinion

Quiz - examples from data (human performance)

1. why its very hard to get an approval for female working visa here in Qatar?

Factual

1. Hi;Can any one tell me a place where i can have a good massage drom philipinies????? yesterday i had a massage in Bio-Bil they charged me 300qr for 01 hour bt it is totally waste... pls advice me if theres any philipinos....

Opinion

1. Hi; Let's speak today about some transportation devices Words: Baas = Bus SayYara = Car Qetar = Train Safeena = Ark Zawraq= Canoe Quiz 1- How can we say "What is the color of Taxi cars in Doha?" in Arabic? – plz use English letters – 2- What is the English Translation of "Aen Maoqef al-Baas?" Have a nice day :)

Socializing

1. Red Light Jumped,Red Light Jumped. Ooopps !!!! I just crossed the Stop line at Crazy signal; But did not crossed the intersection . The camera Lights flashed 2 times and I stood there only. Will it be fined adn how much????????

Factual

1. Qatar CV format? Hey everyone; I'm applying for a job in Qatar; and I'm wondering how to format my CV. I already have a suitable CV for the UK; so please could someone let me know if there's any extra information that I need to add to my CV; to meet the standards in Qatar? Thanks in advance! :)

Opinion

according to our evaluation, human-level accuracy on this data set was about 0.75, which was relatively low

1118 TRAIN (labeled data samples) / **239** DEV / **953** TEST

Data samples given in format:

- **id**
- **label** - F/O/S
- **category** - one of forum categories in form of text (e.g., “Welcome to Qatar”)
- **date**
- **userID**
- **username** - nickname
- **subject** - english text
- **text** - question, english text

sometimes empty text (question body) - question in subject
sometimes not english / mixed / far from correct english language

highly unbalanced / classes distribution changes across sets a lot
e.g., : “Visas and permits” category: 8% in TRAIN, 5% in DEV, 30% in TEST

We tested a few simple text preprocessing and normalization setups:

- urls replacement
- numbers substitution
- custom Qatar dictionary, e.g.:
 - qling >> browsing Qatar forum
 - villagio >> Qatar shopping center
- spell correction with custom character-based CNN language model

...and some others

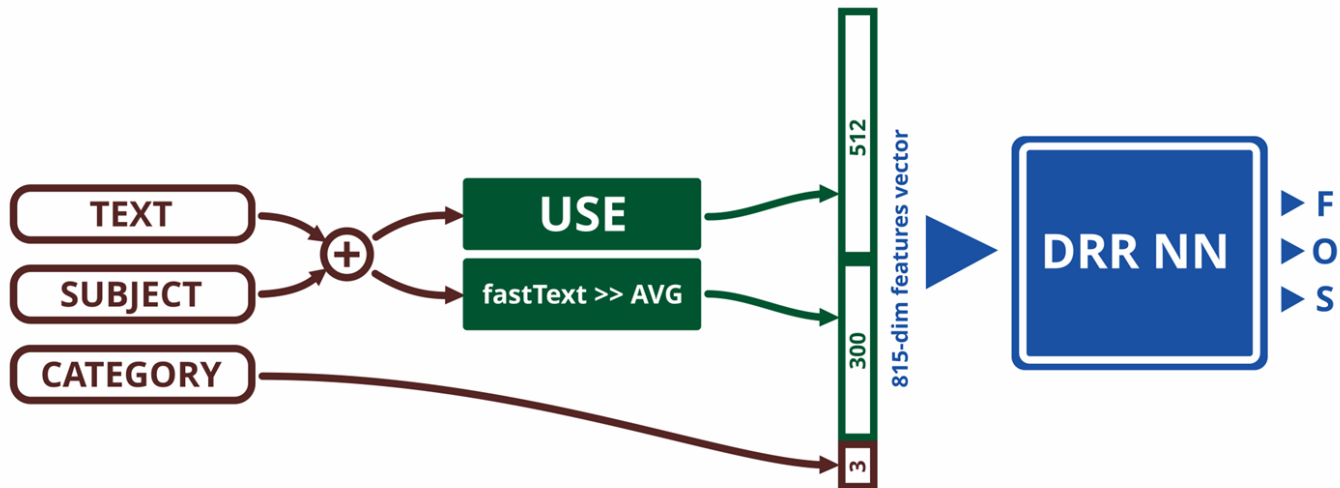
We tested a few simple text preprocessing and normalization setups:

- urls replacement
- numbers substitution
- custom Qatar dictionary, e.g.:
 - qling >> browsing Qatar forum
 - villaggio >> Qatar shopping center
- spell correction with custom character-based CNN language model

...and some others

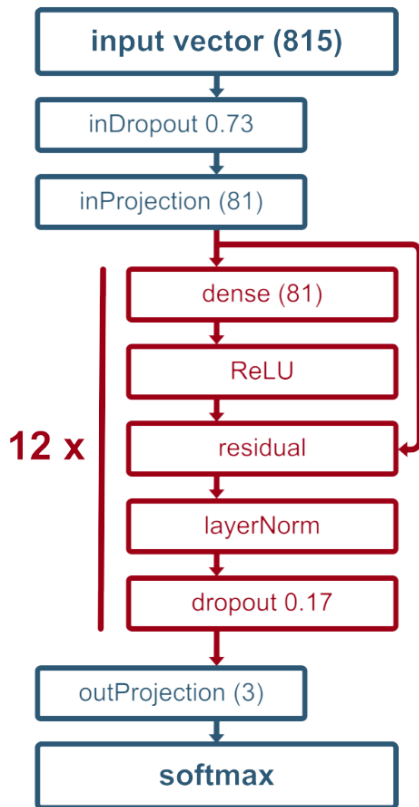
Unfortunately, none of them helped the models achieve improved results, probably preprocessing **deprived data** of significant **features**

System overview



- subject text concatenated with question body text
- USE (Universal Sentence Encoder from Google) 512-dim vector
- word tokenization >> fastText vectors >> averaged >> 300-dim vector
- F/O/S distribution over categories >> 3-dim vector

all concatenated and feed into DRR NN (classifier network)



Deeply Regularized Residual Neural Network

- large input dropout (on 815-dim vector), then projection to 81-dim
- 12 blocks of dense with ReLU, residual, layer norm and dropout
- 3-dim (F/O/S) output projection with softmax probabilities

hyperparameters optimized with random search algorithm

about 148K learnable variables

The problem: **1118** samples for **deep model of 148K** parameters

- we split training data into bigger learning part and smaller validation part
- loss was minimized on the learning part until the accuracy on the validation was increasing
- we used 5-fold cross-validation splits
- to aggregate more knowledge we prepared 4 such splits with different random seeds
- this procedure gave us a total of 20 parts of learning/validation sets
- max num of epochs for training was set to 700, accuracy usually was improving till 300-600
- final prediction with argmax of summarized softmax of 20 models

Contrastive submissions

For the contrastive submissions, our overall idea was to utilize multiple models that were as varied as possible, and combine their outputs.

We used ELMO, BERT, DRR NN and Bag-of-words solutions to build ensemble models that combined submodels output probabilities.

Model	Accuracy
ELMO	0.72
BERT	0.717
Bag-of-words (TF-IDF + SVM)	0.699

models accuracy measured on the DEV set

Results

Primary submission was ranked as the second one among all (21) participants in the official evaluation phase.

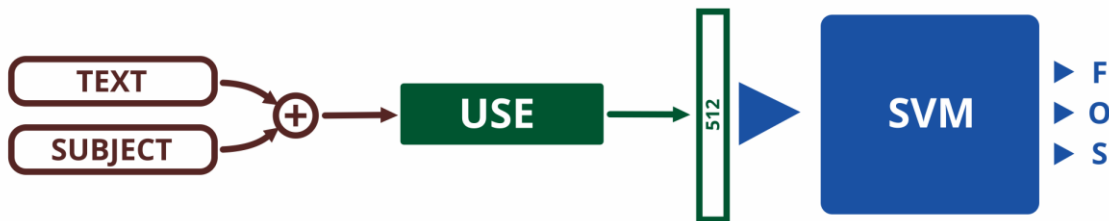
Model	Accuracy	F1	AvgRec
DRR NN (Primary)	0.83	0.72	0.76
Contrastive-1	0.83	0.72	0.76
Contrastive-2	0.81	0.69	0.73
Post-evaluation	0.87	0.77	0.78

results of official evaluation

- **Contrastive-1** used the SVM classifier with linear kernel.
- **Contrastive-2** was designed as a bagging classifier of 10 estimators, each a voting ensemble of logistic regression, random forest and SVM with linear kernel

Conclusions

1. We attribute DRR NN good performance on the small data set to the deep regularization and the advanced training procedure.
2. Usual high performers, such as BERT or ELMO and the ensemble systems could not outperform our primary submission probably due to data overfitting and lack of ability to extract higher-level dependencies from the provided samples.
3. Small data set and distribution differences were main challenges of this competition.
4. SVM classifier (post-evaluation submission) does it better, at least for small data sets.



post-evaluation submission architecture

Thank You



SAMSUNG

© 2017. Samsung R&D Institute Poland. All rights reserved.