Review report of a final thesis

Czech Technical University in Prague **Faculty of Information Technology**

Student:	Šimon Let
Reviewer:	Ing. Jan Motl
Thesis title:	Phishing Email Detection based on Entity Recognition
Branch of the study:	Software Engineering

Date: 7. 6. 2017

Evaluation criterion:	*The evaluation scale: 1 to 5.*
1. Difficulty and other comments on the assignment	*1 = extremely challenging assignment,* *2 = rather difficult assignment,* ***3 = assignment of average difficulty,*** *4 = easier, but still sufficient assignment,* *5 = insufficient assignment*

Criteria description:
Characterize this final thesis in detail and its relationships to previous or current projects. Comment what is difficult about this thesis (in case of a more difficult thesis, you may overlook some shortcomings that you would not in case of an easy assignment, and on the contrary, with an easy assignment those shortcomings should be evaluated more strictly.)

Comments:
The methodology is well known for English language. Hence, the assignment cannot be treated as difficult.
On the other end, converting Natural Language Processing approaches from one language to another is not a simple task.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
2. Fulfilment of the assignment	***1 = assignment fulfilled,*** *2 = assignment fulfilled with minor objections,* *3 = assignment fulfilled with major objections,* *4 = assignment not fulfilled*

Criteria description:
Assess whether the thesis meets the assignment statement. In Comments indicate parts of the assignment that have not been fulfilled, completely or partially, or extensions of the thesis beyond the original assignment. If the assignment was not completely fulfilled, try to assess the importance, impact, and possibly also the reason of the insufficiencies.

Comments:
The amount of work done on improving NER for the domain is noteworthy.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
3. Size of the main written part	***1 = meets the criteria,*** *2 = meets the criteria with minor objections,* *3 = meets the criteria with major objections,* *4 = does not meet the criteria*

Criteria description:
Evaluate the adequacy of the extent of the final thesis, considering its content and the size of the written part, i.e. that all parts of the thesis are rich on information and the text does not contain unnecessary parts.

Comments:
The length of the thesis is appropriate for the content.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
4. Factual and logical level of the thesis	*90 (A)*

Criteria description:
Assess whether the thesis is correct as to the facts or if there are factual errors and inaccuracies. Evaluate further the logical structure of the thesis, links among the chapters, and the comprehensibility of the text for a reader.

Comments:
The justification of the used datasets, NER implementation, type of classifier, evaluation metrics and methodology is well done.

I appreciate that it was acknowledged that NERs may not generalize well to different domains and that the NER was retrained.

I also appreciate that the results are reported for both, Czech and English datasets. Hence, by comparison of the obtained results on well-known English datasets to the results in the literature we can get an idea about the quality of the solution.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
5. Formal level of the thesis	*70 (C)*

Criteria description:
Assess the correctness of formalisms used in the thesis, the typographical and linguistic aspect s, see Dean's Directive No. 14/2015, Article 3.

Comments:
Just in Czech abstract:
- Náše řešení -> Naše řešení
- Vyvinuli jsme rešení -> Vyvinuli jsme řešení
- Řešení dosáhlo -> Řešení dosáhlo
- současně nepřekonaný... -> ...v současnosti nepřekonaný...

In English abstract:
- companies wolrdwide -> many companies worldwide
- Target based -> target-based
- For Named entity recognition we... -> For named entity recognition, we...
- Finally the... -> Finally, the...

Generally:
1) Sentences are frequently missing commas.
2) Sentences sometimes miss a verb. Example:
Supervised learning the main focus of this thesis.

The terms in Czech abstract are not well chosen.
1) "Feature" is commonly translated as "příznak", not "vlastnost".
2) "Named entity recognition" is traditionally translated as "rozpoznávání pojmenovaných entit", not "rozpoznávání entit".
Since the thesis is in English and the terms in English text are overall well used, it is not a big issue. If I had to nitpick, referencing F-measure as "accuracy" can a bit misleading, since "accuracy" can also mean "classification accuracy". And that is a completely different measure than F-measure.

Evaluation criterion: | *The evaluation scale: 0 to 100 points (grade A to F).*

6. Bibliography — *80 (B)*

Criteria description:
Evaluate the student's activity in acquisition and use of studying materials in his thesis. Characterize the choice of the sources. Discuss whether the student used all relevant sources, or whether he tried to solve problems that were already solved. Verify that all elements taken from other sources are properly differentiated from his own results and contributions. Comment if there was a possible violation of the citation ethics and if the bibliographical references are complete and in compliance with citation standards.

Comments:
The text is thoroughly referenced. And the references are relevant.

However, the position of the references varies a lot. Sometimes they are at the beginning of a sentence, sometimes at the end of the sentence. And sometimes they are behind the sentence. Examples:
1) It was successfully used alongside with other methods in [24] and [26].
2) [48] Some features were carefully selected form previous papers other features were added by author himself.
3) The classification features utilize morphological analysis, two-stage prediction, word clustering and gazetteers. [46]

Sometimes the references are not factually correct. For example:
According to [24] Linear regression has the best results out of all compared methods.
But the article describes logistic regression, not linear regression.

Some references are incomplete. For example, reference [36] misses the name of the proceeding (CEET) and page range (131-135).

Evaluation criterion: | *The evaluation scale: 0 to 100 points (grade A to F).*

7. Evaluation of results, publication outputs and awards — *80 (B)*

Criteria description:
Comment on the achieved level of major results of the thesis and indicate whether the main results of the thesis extend published state-of-the-art results and/or bring completely new findings. Assess the quality and functionality of hardware or software solutions. Alternatively, evaluate whether the software or source code that was not created by the student himself was used in accordance with the license terms and copyright. Comment on possible publication output or awards related to the thesis.

Comments:
As the author says: "Observations made in this thesis should be confirmed by testing using bigger dataset or live traffic."

The experience is that the real-world behaviour (from "online" testing) can differ significantly from the lab results (from "offline" testing). Without the online testing, the results are unlikely to be publishable. But it is unreasonable to expect measurements from online testing in a bachelor thesis.

Evaluation criterion: | *No evaluation scale.*

8. Applicability of the results

Criteria description:
Indicate the potential of using the results of the thesis in practice.

Comments:
Based on article "Phishing Email Detection in Czech Language", there is a chance for the deployment.

Evaluation criterion: | *No evaluation scale.*

9. Questions for the defence

Criteria description:
Formulate any question(s) that the student should answer to the committee during the defence (use a bullet list).

Questions:
1) How is it possible that the trained NER has F-measure of 80.37 on 7 classes (types) and 82.11 on 46 classes (subtypes)? Just by chance, we would expect significantly lower F-measure on 46 classes than on 7 classes. Are the reported differences so small because the reported F-measures are micro averages (and not macro averages)?

2) Can you show us a funnel diagram with the estimated proportion of emails from all email traffic, on which the NER features are applicable? The diagram should cover the email categorization from section 1.1 and possible issues with the NER and the target identification.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
10. The overall evaluation	*80 (B)*

Criteria description:
Summarize the parts of the thesis that had major impact on your evaluation. The overall evaluation **does not** have to be the arithmetic mean or any other formula with the values from the previous evaluation criteria 1 to 9.

Comments:
There were many hurdles to pass ranging from small, dirty and unlabelled data to high computational requirements (300GB of RAM...). Still, the author had time to diagnose the reasons behind the initially poor accuracy of the model and satisfactorily address them.

Signature of the reviewer: