

Jan Šnajder, Associate Professor
Faculty of Electrical Engineering and Computing
University of Zagreb

Zagreb, 30 April, 2020

Subject: Filip Boltužić, Master of Science in Computing
– Evaluation of the doctoral thesis –

**Supervisor consent and assessment of
the performed research and the original scientific contribution achieved**

The doctoral thesis “*Computational Methods for Argumentation Mining of Claims in Internet Discussions*” (“*Računalni postupci dubinske argumentativne analize tvrdnji u internetskim raspravama*”) by Filip Boltužić, Master of Science in Computing, is the result of natural language processing research to develop argumentation mining techniques to structure argumentation and detect implicit claims from internet discussions. Argumentation mining provides the argumentative structure of a discussion, which has a broad application scope, including but not limited to social science research interested in critically analysing text, business and political stakeholders attempting to understand public opinion, and education professionals teaching critical thinking skills. Use of machine learning can significantly improve argumentation mining, aid in the detection of implicit claims in text, and provide various additional insights. In particular, structured prediction models, a special case of supervised machine learning techniques trained to predict structured objects, have been employed to model semantic relationships in text to improve claim detection.

The thesis proposes using ontologies to solve claim structuring. Ontologies are used to model specific domain concepts and claim patterns in order to conduct argumentation analysis of internet discussions. A framework that involves building a domain specific ontology, claim detection, and claim structuring is proposed to structure claims from the text of the discussion.

The doctoral thesis is structured as follows. The first chapter (“1. *Introduction*”) motivates the proposed research, describes the research goals, and provides an overview of the thesis. The rest of the thesis can be conceptually divided into three parts. The first part (chapters two and three) introduces the area of argumentation mining, discusses relevant related work in argumentation mining, and outlines machine learning methods applied in argumentation mining with an emphasis on structured prediction. The second part (chapters four, five, and six) describes preliminary studies, which applied unstructured approaches based on natural language processing to solve argumentation mining problems. In the third part (chapters seven, eight, nine, and ten) ontology-based methods are proposed for claim structuring. The proposed methods are setup as part of framework to detect and structure claims. Finally, an example argumentative analysis using the proposed framework is demonstrated.

The second chapter (“2. *Argumentation Mining*”) provides a brief overview of argumentation mining. The chapter also introduces relevant related work from argumentation mining. Argumentation mining problems are categorized into argumentative component detection and argumentative component structuring. Past approaches to each of the two sets of problems are described. The end of the chapter outlines some of the shortcomings of previously used argumentation mining methods. The third chapter (“3. *Methods and Tools for Argumentation Mining of Claims*”) reviews supervised machine learning, unsupervised machine learning, and structured prediction methods and concepts. A short review of selected topics from natural language processing and ontologies is also provided.

The fourth chapter (“4. *Claim Clustering*”) introduces the problem of prominent claim extraction from internet discussions. Unsupervised machine learning methods are used to provide solutions to the problem. The experiments are carried out on a dataset of internet discussion texts which contain comments and their respective prominent claims. A hierarchical clustering method using semantic similarity as a distance measure is proposed. The centroids of the resulting clusters are then named prominent claims. The hierarchical clustering results are evaluated quantitatively and qualitatively. The evaluation results are thoroughly discussed, based on which future work is proposed.

The fifth chapter (“5. *Prominent Claim Identification*”) defines the problem of prominent claim identification. A corpus of comments from two internet discussions (“*Gay rights*” and “*Under God in pledge*”) is introduced. The dataset consists of pairs of comments and prominent claims the comments refer to. The problem of prominent claim identification is then defined as determining the relationship between the comment and prominent claim. Semantic similarity and textual entailment features are used as input to supervised machine learning methods. The models are experimentally evaluated on the newly introduced corpus. Since the proposed methods are domain independent, experiments across different discussions are conducted. A thorough error analysis of the proposed models is made, resulting in proposed future work guidelines.

The sixth chapter (“6. *Deriving Implicit Claims*”) incorporates insight from the previous two chapters to define the problem of implicit claim detection from internet discussions. The dataset described in chapter four is extended by annotating implicit claims between comments and prominent claims. Implicit claims make the logical reasoning between the comment and prominent claims valid. Another supervised machine learning model to solve the prominent claim detection problem is proposed, with the difference of incorporating information from implicit claims. Using implicit claim information demonstrated significant improvements, so preliminary models to automatically derive implicit claims is proposed. The preliminary model does not perform well, only further highlighting drawbacks of unstructured approaches to argumentation mining.

Structured approaches to argumentation mining are introduced in the chapter seven (“7. *Claim Segmentation*”). Structuring claims is decomposed into two steps: claim segmentation and structuring. The problem of claim segmentation is defined analogously to conceptually related natural language processing problems as a sequence classification problem. Several structured prediction models to solve sequence classification are proposed. The dataset introduced in chapter four is again extended by annotating segments in comments. The proposed models are experimentally evaluated, which leads to the conclusion that structured prediction methods improve claim segmentation.

Chapter eight (“8. *Formalizing Claims*”) introduces structured claims. Structured claims abstract claim content just enough to reduce the problem of language variation, yet specific and complex enough to attain the argumentative gist of claims. A two-level ontology is then used to formalize and structure claims. An annotation study is carried to structure claims from the “*Marijuana legalization*” topic.

To complete the framework of claim segmentation and structuring, chapter nine (“9. *Claim Structuring*”) studies the claim structuring problem. Several structured prediction models to solve claim structuring are proposed. Since related work has shown that multitask learning can often yield performance boosts when done on related tasks, a joint model that segments and structures claims is proposed. Experimental evaluation is conducted, with structured prediction models demonstrating the best performance.

Chapter ten (“10. *Analysis using Formalized Claims*”) shows an example of an argumentation mining analysis of claims. In contrast to the approach shown in chapter six to retrieve implicit claims, a logic-based approach is demonstrated. Applications of structured claims are demonstrated on the “*Marijuana legalization*” topic by identifying most frequent claims and clustering claim authors based on their shared claims.

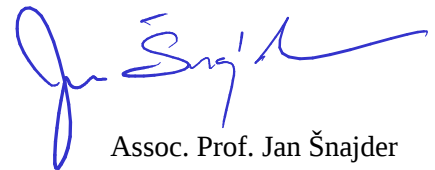
The final chapter, chapter eleven (“11. *Conclusion*”) concludes and provides future work guidelines.

I, Jan Šnajder, am of the opinion that Filip Boltužić, Master of Science in Computing, has conducted thorough research and has adequately presented his findings through his doctoral thesis. The results of his research have

been submitted for peer review in two international journals indexed by the *Science Citation Index Expanded* and published in four international conference papers. The research contribution is as follows:

1. Modeling an online discussion using an two-level ontology, where the first level contains domain knowledge and the second models claim patterns, aiming to fully structure online discussions;
2. Supervised machine learning method for claim detection and claim structuring;
3. Framework and support for online discussion analysis involving claim detection, structuring, and analysis based on a comparison of claims from all discussion participants.

In conclusion, I give my consent to submit the doctoral thesis of Filip Boltužić, Master of Science in Computing, for evaluation.



Assoc. Prof. Jan Šnajder