# Argument Mining in Wikipedia Articles

**Ben Rothschild**
bnroths@uchicago.edu

**Julia Zhou**
juliazhou@uchicago.edu

## Abstract

## 1 Introduction

Argumentation mining aims to identify structured argument data from unstructured text. Applications of argumentation mining are vast and include improving information retrieval from long text, summarizing arguments in legal texts, scientific writing, and news articles or accessing student's command of subject knowledge in essay assignments. Recently argumentation mining has also been used to develop technologies that can assist humans to debate and reason.

The task of identifying arguments is difficult though because of the various forms arguments can take in relation to a document or topic. Part of argumentation mining is also determining if an argument is Pro vs. Con a specific topic. This is called stance identification and is a difficult NLP task because the subtle nature language can play in taking a stance that often relies on context and subtle word choices or stance development within an argument. Current approaches are engineered to address specific domains, for example a specific model might be built just to analyze claims in court documents using attributes specific to court documents and legal vocabulary. However, we believe that argumentative sentences are often characterized by common rhetorical structures, independently of the domain and we propose to explore a method that exploits structured parsing information to detect claims without resorting to topic specific information.

An example of this would be taking a topic like the sale of violent video games harms minors and a wikipedia article about the Video game content rating system and identifying if a specific sentence or entire text of the article has a Pro or Con stance towards the topic. The article includes the sentence Exposure to violent video games causes at least a temporary increase in aggression and this exposure correlates with aggression in the real world which should be labeled as a PRO stance. There are two main tasks in this problem: - Identifying claims in text, this will be done through an entity resolution and scoring task using SVM - Identifying the stance of a claim against a topic, this will be done similar to a sentiment classification task By combining these tasks we will be able to tell what sentences in a long text support and oppose a claim.

## 2 Related Work

LIT REVIEW ANYTHING? (good examples in Roy Bar-Haim paper

## 3 Data

The dataset we are using is the Claim Stance Dataset from IBM Debater project which can be accessed here http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml. It contains 2,394 labeled claims for 55 topics that are pulled from 1,065 wikipedia articles. For each article we are given the following data points:

- Full text from Wikipedia (tex)

- Topic Target (text)

- Claim Text (text)

- Claim Start Index (integer)

- Claim End Index (integer)

- Stance (Pro or Con)

The dataset was created by first looking at the list of controversial issues `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues` Wikipedia identifies to find a subset of 56 topics where there was a clear two-sided debate. From these topics 546 articles were identified that night have claims in them. Annotators then read these articles and identified claims and stances within the articles. An example observation in the dataset would be as follows:

**Full Text** 44,000 words plain text wikipedia article

**Topic Target** the sale of violent video games to minors

**Claim Text** they increase the violent tendencies among youth

**Claim Start Index** 8119

**Claim End Index** 8167

**Stance** PRO

A full description of the data generation process can be found here (Toledo-Ronen et al., 2016)

For each article, the algorithm identifies claims and stance (Pro/Con) towards the topic. Additional fine-grained annotations such as topic target, topic sentiment towards its target, claim target, claim sentiment towards its target, and the relation between the targets are also included based on the semantic model of Bar-Haim et al

## 4  Method Overview

We divided this research problem into two parts, claim identification and stance identification.

### 4.1  Claim Identification

### 4.2  Stance Identification

## 5  Results and Analysis

## 6  Conclusion

## References

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123.