# CRST: A Claim Retrieval System in Twitter

**[1]Wenjia Ma, [1]Wenhan Chao, [2]Zhunchen Luo†, [1]Xin Jiang**
[1]School of Computer Science and Engineering, Beihang University
[2]Information Research Center of Military Science
PLA Academy of Military Science
[1,2]Beijing, China
[1]{mawenjia, chaowenhan, xinjiang}@buaa.edu.cn
[2]zhunchenluo@gmail.com

## Abstract

For controversial topics, collecting argumentation-containing tweets which tend to be more convincing will help researchers analyze public opinions. Meanwhile, claim is the heart of argumentation. Hence, we present the first real-time claim retrieval system CRST that retrieves tweets containing claims for a given topic from Twitter. We propose a claim-oriented ranking module which can be divided into the offline topic-independent learning to rank model and the online topic-dependent lexicon model. Our system outperforms previous claim retrieval system and argument mining system. Moreover, the claim-oriented ranking module can be easily adapted to new topics without any manual process or external information, guaranteeing the practicability of our system.

## 1 Introduction

When users search controversial topics in Twitter, they often tend to find some persuasive tweets. Argumentation is known as the most convincing structure, which usually consists of claim and evidence (Toulmin, 1958; Palau and Moens, 2009). However, only when the claim confirmed, can the evidence make sense. To help users swiftly obtain many pre-eminent claims about the query topic from Twitter, we propose CRST, a system that can automatically retrieve claim-oriented tweets.

Given a topic, our task aims to retrieve a list of claim-oriented tweets. We assume a claim-oriented tweet should meet three criteria: (1) the tweet should be topic-related; (2) the tweet clearly supports or opposes the topic; (3) the tweet provides an arguable reason for its stance. For example, "*@mmfa Abortion is not a choice, abortion is the killing of an innocent life.@owillis*" is a tweet related to the topic of "*abortion*". Moreover, it strongly opposes abortion and contains an arguable reason, "`abortion is the killing of an innocent life`". Therefore, it is a claim-oriented tweet.

To the best of our knowledge, this is the first attempt of claim retrieval in Twitter. Most existing works on argument mining in Twitter concentrate on detecting the evidence types (Dusmanu et al., 2017). And the claim retrieval task on documents was first introduced by Roitman et al. (2016). However, due to the short tweet content and specific conventions in Twitter as well as the ambiguous claims made by tweeters, our task is harder than claim retrieval in documents.

CRST integrates search and re-ranking modules to (i) find topic-related tweets, and (ii) rank by the degrees of containing claim. The core NLP part of our system is the claim-oriented ranking module (see Section 2.2 for detail). It can be divided into the offline topic-independent learning to rank model and the online topic-dependent lexicon model. Considering (1) some conventions in Twitter structure tweets and this structuring can be a valuable hint for searching claim-oriented tweets; (2) some claims may be expressed in a general pattern; we use a learning-to-rank framework to integrate Twitter structure information and some general claim pattern features to build an offline topic-independent ranking model. In addition, claims can not exist without topic, so we introduce the topic information to our claim-oriented ranking module. To be more specific, we generate a topic-dependent claim-oriented lexicon online to

† Corresponding author.

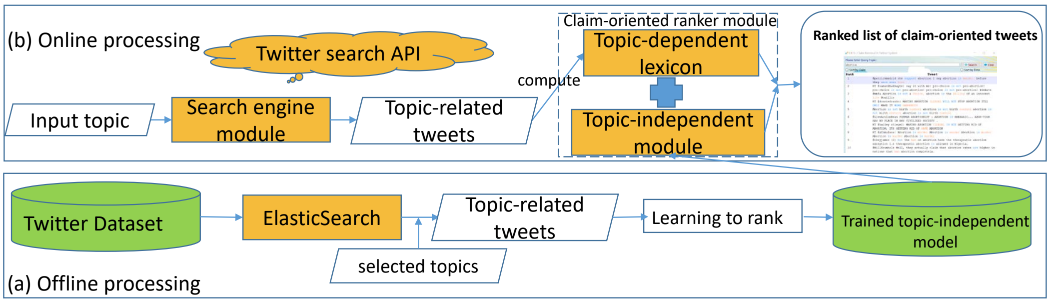


Figure 1: Overview of our system for retrieving claim-oriented tweets.

further elevate the retrieval performance. Experimental results show that our system outperforms other systems on similar tasks.

## 2 Claim Retrieval System

### 2.1 System Overview

An overview of our system is shown in Figure 1. We first perform offline steps to process data and to train the topic-independent model (Subfigure a). The online system is illustrated in Subfigure b. In the remainder of this section, we briefly discuss these steps.

**Offline Processing** In order to build the system, we crawl and index about 60 million English tweets using the Twitter API in 2016. Using these tweets we implement a search engine based on ElasticSearch. Given a query topic, the search engine would present a list of relevant tweets ranked based on the BM25 score. We construct a annotated dataset by searching some selected topics on the search engine (see Section 4.1 for detail). Then, we train a learning-to-rank framework which integrates different kinds of topic-independent features as a topic-independent model.

**Online Processing** When a user gives a new query topic $q$, the system performs the following three steps on the fly: (i) *Retrieving related tweets with a real time Twitter search API*, where Tweepy[1] is invoked to retrieve the top-n tweets that are most related to $q$; (ii) *Ranking the tweets*, where we automatically construct topic-dependent claim-oriented lexicons online and combine it with the offline trained topic-independent module as our ***Claim-oriented Ranking Module*** (elaborate in Section 2.2). (iii) *Visualizing the results*, where the visualization module presents the re-ranked tweets to the user within an interactive graphical interface.

### 2.2 Claim-oriented Ranking Module

By and large, our retrieval model is a learning-to-rank framework which integrates topic-independent features. Additionally, we use topic-dependent claim-oriented lexicons to further elevate the retrieval performance. Given a query topic $q$, a list of related tweets $T$ from the Twitter dataset $D$ is calculated as $T = Relevant(D, q)$[2]. The final claim-oriented score function of a tweet $t$ is defined as $FinalScore(t, q) = LTR(T, t) + \lambda ScoreLex(t, Lex_q)$, where $LTR(T, t)$ is a pairwise learning to rank method[3] and $ScoreLex(t, Lex_q)$ is a function[4] using a claim-oriented lexicon $Lex_q$ to construct an claim-oriented score for each tweet $t$. $\lambda$ is a hyper parameter obtained through training. And we will elaborate them in the following part.

**Topic-independent Module** We use learning to rank framework to build our topic-independent model. Learning to rank is a data driven approach that effectively incorporates a bag of features into the retrieval

[1] `http://www.tweepy.org/`

[2] We used Okapi BM25 as a the *Relevant* function.

[3] We use rankSVM (`http://www.cs.cornell.edu/people/tj/svm\_light/svm\_rank.html`) to train our ranking model.

[4] We estimate the claim-oriented score of each tweet by calculating the average claim-oriented score over certain terms.

process. We use a linear kernel rankSVM for training and report results for the best setting of parameters. There are three kinds of topic-independent features that we used. (i) *Twitter structural features* refer to conventions that are only used in Twitter, such as "***#***", "@", "***RT*** @", ***URL*** and combinations of these conventions. Meanwhile, we also consider whether a tweet is a ***reply*** as a feature. (ii) *Author social features* contain ***followers***, ***friends*** and ***status*** information, which also play an important role in social media mining work (Luo et al., 2012; Luo et al., 2015). (iii) $GenLex$ is a topic-independent claim-oriented lexicon. We compute the information gain score of each word in tweets which contained in a labeled dataset. Then we choose the top-N scored words to construct $GenLex$. These words, such as "because", "but", "will", are topic-independent words and can reveal general patterns of claim-oriented tweets.

**Topic-dependent Claim-oriented Lexicon** Since it is impossible to train a supervised model for every topic, we adopt the $ScoreLex(t, Lex_q)$ (mentioned above) by considering it as a topic-dependent problem. The topic-dependent claim-oriented lexicon $Lex_q$ is constructed by $MakeLex(GenLex, T)$, where we measure topic-dependent claim-oriented score of each word $w$ in $T$, by calculating the co-occurrence frequency of $w$ with words in $GenLex$. Finally, the high scored words will be used to construct a claim-oriented lexicon that refers to query topic $q$. For example, about topic "abortion", the words in $Lex_q$ are "kill", "murder", "dangerous", etc.

### 2.3 System Usage

Our system is shown in Figure 2. The interface of our system is implemented using PyQt5. After inputting the query topic $q$, users can choose to use a claim-oriented score order or time sort. In the former model, our system automatically highlights the words in $GenLex$ to blue, and the words in $Lex_q$ to yellow. This can be seen as a basis for tweets to be considered as containing claims. Additionally, our system allow users viewing details by clicking the tweet.

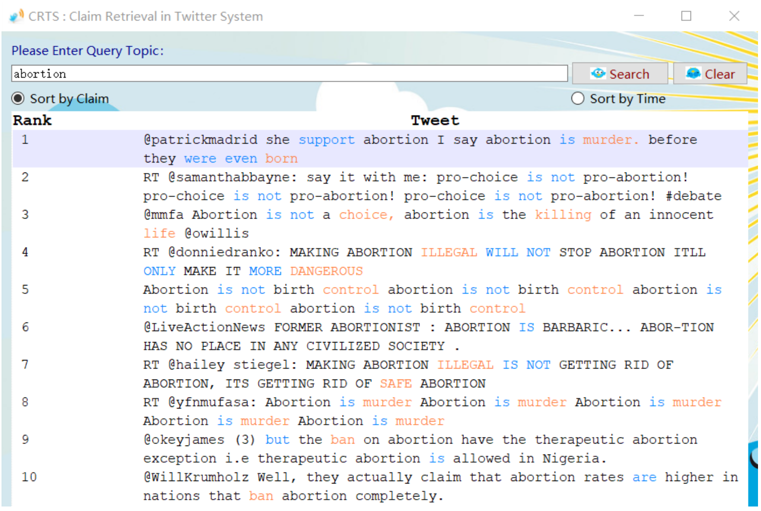

Figure 2: Homepage screenshot of our system.

## 3 Experimental Result and Case Study

### 3.1 Evaluation

To evaluate our claim retrieval model, we construct a English Twitter dataset which consists of 2520 tweets from 30 controversial topics[5]. And 586 tweets are identified as containing claims. We use 10-fold cross validation for evaluation, and use MAP as metric.

We use $WikiClaim$ and $TwitArgument$ as baselines. We adopt the features which are used for retrieving claims in wikipedia documents in Roitman et al. (2016), and name it $WikiClaim$.

[5]We choose topics from `https://www.procon.org/`

| Rank | Tweet | Reply |
|---|---|---|
| 1 | @patrickmadrid she support abortion I say abortion is murder. before they were even born | 1 |
| 2 | RT @samanthabbayne: say it with me: pro-choice is not pro-abortion! pro-choice is not pro-abortion! pro-choice is not pro-abortion! #debate | 0 |
| 3 | @mmfa Abortion is not a choice, abortion is the killing of an innocent life.@owillis | 1 |
| 4 | RT @donniedranko: MAKING ABORTION ILLEGAL WILL NOT STOP ABORTION ITLL ONLY MAKE IT MORE DANGEROUS | 0 |
| 5 | Abortion is not birth control abortion is not birth control abortion is not birth control abortion is not birth control | 0 |
| 6 | @LiveActionNews FORMER ABORTIONIST : ABORTION IS BARBARIC... ABORTION HAS NO PLACE IN ANY CIVILIZED SOCIETY . | 1 |
| 7 | RT @hailey_stiegel: MAKING ABORTION ILLEGAL IS NOT GETTING RID OF ABORTION, ITS GETTING RID OF SAFE ABORTION | 0 |
| 8 | RT @yfnmufasa: Abortion is murder Abortion is murder Abortion is murder Abortion is murder Abortion is murder | 0 |
| 9 | @okeyjames (3) but the ban on abortion have the therapeutic abortion exception i.e therapeutic abortion is allowed in Nigeria. | 1 |
| 10 | @WillKrumholz Well, they actually claim that abortion rates are higher in nations that ban abortion completely. | 1 |

Table 2: Examples of our system for the querying "abortion".

We also adopt the features which are used for argument identification tasks in Twitter in Theodosis Goudas and Karkaletsis (2015), and name it $TwitArgument$. Considering topic related factor, we combine BM25 with them. As shown in Table 1, our best model ($Best$) which use both learning to rank framework to integrate topic-independent features and topic-dependent claim-oriented lexicon outperforms the baselines significantly.

| Methods | MAP |
|---|---|
| $WikiClaim + BM25$ | 0.291 |
| $TwitArgument + BM25$ | 0.328 |
| $Best$ | **0.585** |

Table 1: Results of Baselines and our best model. $Best$ significantly better than baselines (for $p < 0.01$).

### 3.2 Case Study

In this section, we demonstrate a scenario of retrieving a query to prove the effectiveness of our system. Table 2 shows the top 10 retrieval results returned by our system when searching for "abortion".

As shown in Table 2, we can figure out that the tweets containing claims are in the top rank such as 1, 2, 3, 4, 5, 6, 7, 8, 10. From these tweet, we can find that many claim-oriented tweets contains a re-tweet feature "***RT*** @", it is very possible because of the high forward frequency of valuable claim. As for the "***reply***" features appear many times, it may be because the argumentation always occurs during the discuss or quarrel. At the same time, some structural features like ***URL*** which is often contained in news or an advertisement rarely appear. In addition, these claim-oriented tweets contain words, like "kill", "life", "murder", which show our model can capture the topic-dependent claim-oriented information.

## 4 Conclusion

We present the first system that supports users to retrieve claim-containing tweets about controversial topics in Twitter. We train a rankSVM for our learning-to-rank framework and the topic-dependent lexicon is constructed using unlabeled topic-related tweets. Hence, our model can be easily adapted to new emerging topics in Twitter. In addition, our system let the user intuitively obtain the claims, which is certainly helpful in the development of public opinion research. The experimental results show that our system outperforms the previous state-of-art document claim retrieval system and Twitter argument mining system.

## Acknowledgements

We appreciate the comments from anonymous reviewers. This work is supported by National Key Research and Development Program of China (Grant No. 2017YFB1402400) and National Natural Science Foundation of China (No. 61602490).

## References

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2317–2322.

Zhunchen Luo, Miles Osborne, and Ting Wang. 2012. Opinion retrieval in twitter. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

Zhunchen Luo, Miles Osborne, and Ting Wang. 2015. An effective approach to tweets opinion retrieval. *World Wide Web*, 18(3):545–566.

Raquel Mochales Palau and Marie Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *International Conference on Artificial Intelligence and Law*, pages 98–107.

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. 2016. On the retrieval of wikipedia articles containing claims on controversial topics. In *International Conference Companion on World Wide Web*, pages 991–996.

Georgios Petasis Theodosis Goudas, Christos Louizos and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and social media. In *International Journal on Artificial Intelligence Tools*, pages 287–299.

Stephen Toulmin. 1958. The uses of argument. *Ethics*, 10(1):251–252.