ng.aux

Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Human-Computer Interaction

Mining Arguments from Experts

Master's Thesis

Fan Fan Matriculation Number 118043
b. 15.09.1994 in GANSU

First Referee: Prof. Dr. Benno Stein
Second Referee: PD Dr. Andreas Jakoby

Submission date: 25.08.2020

Abstract

This thesis represents a system to automatically identify the stances of arguments from experts towards controversial topics by using the background knowledge and provide a solution for people to efficiently explore these arguments with different stances and get a trustworthy overview about the topic they are interested in. Specifically, there are two components introduced in the thesis: (1) Stance Knowledge Graph (SKG): Following distance supervision techniques, we constructed the SKG which encodes 149,683 stances from 82,501 experts to 170 controversial topics and 604 high-frequent topic correlation patterns in the stance perspective; (2) Argument from Experts Mining: We proposed an SKG-based model to identify the stances of 26,524 mined arguments from experts towards specific topics on a large set of quotations. In the end, we integrated our work into Args.me (an argument search engine)[43] to provide a solution for people to efficiently explore ranked arguments from the experts about the topic they have queried.

Acknowledgements

I would like to express my deepest appreciation to Dr. Khalid Al-Khatib and Yamen Ajjour. This project would have been impossible without their patience, support and immense knowledge. I cannot imagine to have better supervisors for my thesis. Their supervision not only encouraged and supported me to finish my thesis, but also motivated me to improve myself and become a better person. I am very grateful that they contributed their time and energy to guide and help me regularly throughout the duration of this thesis to solve the difficulties I have experienced and make this work possible.

I would like to extend my sincere thanks to my friends as well, Ren He and Franz Hack, for their selfless and continuous support and inspiration. With their help, I could keep challenging myself to jump out of my comfort zone without concerns and have a meaningful life in Germany.

In the end, special thanks to my parents, who gives me full support to chase my academic goal in Germany and enrich my life experience. Without their unselfish support, all of these probably will not happen.

Thanks for all your encouragement!

Contents

List of Figures 6

List of Tables 6

1 Introduction 7
1.1 Stance Knowledge Graph (SKG) 8
1.2 Argument from Expert Mining 10
1.3 Thesis Contribution 12
1.4 Thesis Structure 12

2 Related Work 13
2.1 Ethos Mining 13
2.2 Argument Scheme 14
2.3 Argument Mining 14
2.4 Stance Classification 15
2.5 Quotation Detection 18

3 Stance Knowledge Graph 19
3.1 Corpus Preparation 19
3.1.1 Seed Corpus 19
3.1.2 Topic Corpus 19
3.1.3 Expert Corpus 20
3.1.4 Expert Stance Evidence (ESE) Corpus 21
3.1.5 Corpus Evaluation 23
3.1.5.1 Approach 23
3.1.5.2 Discussion 23
3.2 ESE Stance Classifier 24
3.2.1 Features Extraction 24
3.2.2 Models Building 25
3.2.3 Experiment Setup 26
3.2.4 Discussion 27
3.2.5 Result 29
3.3 SKG Construction 29
3.3.1 Topic-Topic Level 30
3.3.2 Topic-Expert Level 32
3.4 SKG Visualization 33
3.4.1 Topic-Topic Level Visualization 33
3.4.2 Topic-Expert Level Visualization 35
3.5 SKG Evaluation 36
3.5.1 Approach 36
3.5.2 Discussion 37

4 Argument From Experts Mining 39
4.1 SKG-based Model 39
4.1.1 Operating Principle 39
4.1.2 Classification Methods 39
4.2 Argument Retrieval 41
4.2.1 Arguments from Quotes-500k 42

4.2.2 Arguments from English Gigaword 42
4.2.3 Arguments from Twitter 42
4.3 Result 43
4.4 Evaluation 43
4.4.1 Approach 44
4.4.2 Discussion 44
4.5 Application 45
4.5.1 Quote View Module 45
4.5.1.1 Interface Layout 45
4.5.1.2 Interface Design 45
4.5.2 Ranking Method 46

5 Conclusion and Future Work 48

References 51

Statutory Declaration 55

List of Figures

1 Illustrates the key steps of the approach 8
2 Illustrates the SKG modeling 10
3 Illustrates the ethos-evidence mining and stance classification 11
4 Quote view in Args.me 12
5 Occupation distribution 21
6 Graph Building illustration 30
7 Stance Knowledge Graph visualization 34
8 Topic-Topic Level visualization with interactions 35
9 Topic-Expert Level visualization with interactions 36
10 Quote View 45
11 Quote Card 46

List of Tables

1 Examples from the topic corpus illustrating the scheme 20
2 Examples from the expert corpus illustrating the scheme 21
3 Examples from the ESE corpus illustrating the scheme 23
4 ESE corpus evaluation 24
5 Distribution of the data for the stance classifier building 26
6 Results for each machine learning model on the test dataset 27
7 Results for traditional neural network models on the test dataset 28
8 Results for BERT model on the test dataset 28
9 Results of the cross-topic and cross-expert with BERT Model 29
10 Topic-Expert datasets 32
11 Stance knowledge graph sample 33
12 SKG evaluation 38
13 Mined arguments statistics 43
14 SKG-based model evaluation 44

1 Introduction

On the web, arguments are regularly shared among people in different platforms such as debate portals, social media platforms, etc. These platforms help to explore arguments for various controversial issues online and hence support people in decision making. However, getting an overview of the arguments related to controversial topics is time-consuming; people need to explore different platforms and try to search for those arguments that match their need. As a solution, [43] introduces an argument search engine (Args.me) to index and retrieve arguments from five debate portals. Despite the importance of such an engine, we observe that it mainly covers arguments written by ordinary people. Compared to the arguments from *experts*, those who have a deeper understanding in their professional fields, the arguments from ordinary people are more likely to contain misinformation and invalid arguments that can misguide people's conceptions. Moreover, arguments from an expert are known to have higher credibility, represented as the ethos mode of persuasion in Aristotle's rhetoric theory, in addition to logos (the sound arguments) and pathos (the audience's feelings triggered by the speaker), as introduced in [18]. Ethos, which is related to the character of the argument holder, plays an important role in influencing people's beliefs, attitudes and actions[13]. For example, the study of [20] indicates that celebrity endorsements can make unpopular statements more palatable while increasing the level of agreement with already popular opinions.

In the computational argumentation area, where researchers address the mining, assessment, and generation of arguments, the focus is mainly on the logos perspective of arguments, while arguments from experts have been studied scarcely[14].

To fill this gap, this thesis concerns the task of mining arguments from experts in order to integrate them in the current argument search engines, leading to a powerful tool for people to acquire a trustworthy overview of controversial topics.

The task of mining arguments from experts can be tackled in different ways. In this thesis, the task is addressed as follows: starting from a predefined set of controversial topics, we find experts for these topics along with their stances towards the topics, and then we mine the web for arguments that address the controversial topics and written by those experts. More specifically, we approach this task by (1) constructing a stance knowledge graph (SKG), that represents the stances of many experts on various controversial topics, following distance supervision techniques, and (2) mining arguments from experts, identifying their stances using the SKG-based model, and integrating them into Args.me to provide a solution for people to explore arguments from experts for the topic they are interested in.

Figure 1 shows an example of the main steps of our approach. First, we extracted sentences including the specific expert and topic related information from Wikipedia, named expert stance evidence (ESE). For each ESE, we utilized a trained ESE stance classifier for stance classification. Next, we aggregated the predicted stances of these ESEs to conclude this expert's stance towards the topic, which will be encoded into the stance knowledge graph (SKG) as the background knowledge. After that, we mined the arguments from this expert with respect to that topic and used the background knowledge to identify the stances of these arguments. In the end, these mined arguments would be integrated into an argument search engine (Args.me). In the following, we describe those steps in detail.

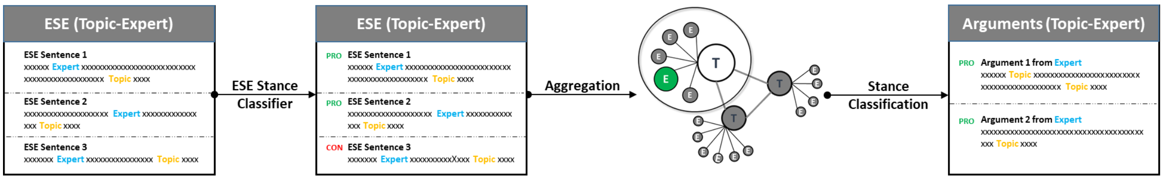


Figure 1: Illustrates the key steps of the approach

1.1 Stance Knowledge Graph (SKG)

Given a natural language text as an input, the stance classification is the task of identifying whether the input text is supportive, opposed, or neutral towards a target entity or topic[33]. Various techniques and approaches have been explored and proposed for this task.

In this thesis, inspired by the Expert Stance Graph introduced in [40], we propose a new large-scale knowledge graph: Stance Knowledge Graph (SKG), which encodes not only the expert' stances to controversial topics but also topic correlation patterns in the stance perspective (see Section 3). Thus, the graph is supposed to indicate experts' ideologies as well as reveal human ideological patterns. This knowledge graph will be utilized as the background knowledge to identify the stances encoded in the *arguments from experts* regarding controversial topics.

For constructing the SKG, we do the following steps:

Corpus Preparation: To determine the expert and topic domains of the SKG, we created two corpora: *topic-corpus* and *expert-corpus*,. The former corpus encodes 170 two-sided controversial topics with their logical correlations and terms using the Debate Topic Expansion dataset[7], WordNet[15] and the Claim Stance dataset (the base of the Expert Stance Graph)[40]. As for the expert-corpus, it contains 446,166 experts: 66,086 from the Expert Stance Graph and 380,080 human entities from the Wikidata list of people.

Using the created corpora, we generated the *Expert Stance Evidence (ESE) corpus* using a 5-step pipeline: (1) converting Wikipedia dumps to structured data, (2) conducting co-reference analysis and segmentation, (3) mining sentences related to the topics and experts in the corresponding corpus, (4) identifying experts in the mined sentences via URL check, and (5) annotating each sentence's stance according to the expert's stance towards the topic encoded in the Expert Stance graph.

The ESE corpus comprises information in the sentence level from Wikipedia including specific experts and topics with annotated stances based on the Expert Stance Graph. According to the stance annotation result, the ESE corpus can be divided into two sets, namely stance-ESE and nonstance-ESE. The former dataset has 46,248 sentences where the experts' stances to the topics are directly identified in the Expert Stance Graph or the stance can be deduced based on the assumption that the expert's stances should be the same to the consistently related topics and be the opposite to the contrastively related topic (the topics' logic relations are encoded in the topic-corpus). The nonstance-ESE dataset contains 130,850 sentences where the experts' stances can be neither identified nor deduced from the Expert Stance Graph.

ESE Stance Classifier: Distant supervision techniques were utilized to construct the SKG. In particular, we trained a BERT-based model (Bidirectional Encoder Representations from Transformers) on the stance-ESE dataset to learn how to classify the given expert's stance towards the given topic by extracting the relation between them shown in the ESE from Wikipedia. The trained model with leveraging the expert and topic related information outperforms the baseline (i.e., SVM, CNN, etc.) with 0.97 F1-macro for the within-domain test as well as 0.59 for cross-topic domain and 0.74 for the cross-expert domain. The adopted model was applied to the nonstance-ESE dataset to predict the stances with the corresponding possibilities of both stances as the confidence scores.

SKG Construction: In the end, based on the Expert Stance Graph and ESE-corpus with the annotated or predicted stances as well as the corresponding aggregation methods, we construct the SKG into two levels, topic-topic level and topic-expert level. The former level encodes 604 high-frequent topic correlation patterns (positive or negative) in the stance perspective, whereas the latter level encodes 149,683 stances (Pro or Con) from 82,501 experts towards 170 topics. Additionally, we implemented the arc diagram visualization technology as well as the necessary interaction techniques (e.g. click, hover, etc.) to represent the built SKG. Fig 2 illustrates the SKG modeling procedure.

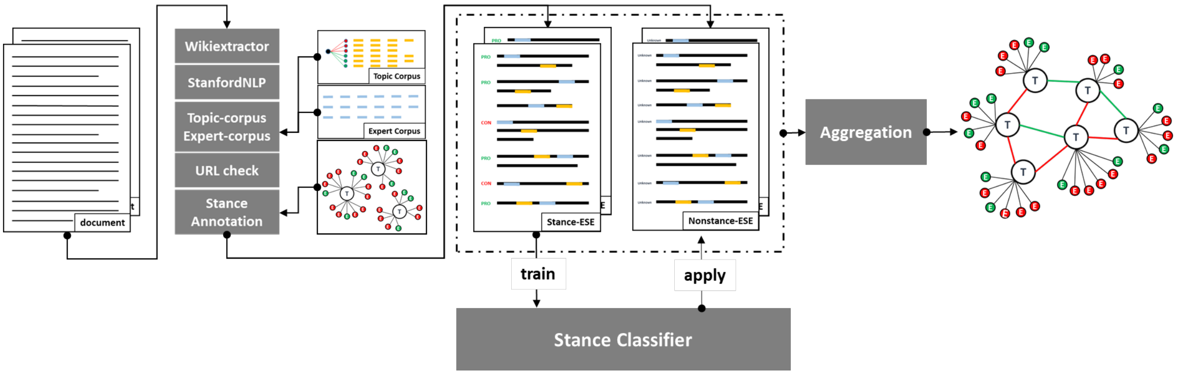


Figure 2: Illustrates the SKG modeling

1.2 Argument from Expert Mining

The procedure for mining arguments from experts with classified stances using the SKG-based model is illustrated in Fig 3.

Argument Retrieval: As a source of arguments from experts, we explore the experts' quotations and tweets. We hypothesis that the quotations and tweets written by a specific expert and including one or more controversial topics are very likely to be arguments.

In specific, we mined expert's quotations from 'quote' datasets such as Quotes-500k[32] and from news datasets such as English Gigaword[16]. For Quotes-500k dataset, which includes cleaned structured quotes, we only filter the quotes identifying those which includes one of our experts and controversial topics listed in the expert and topic corpora. Regarding English Gigaword, which contains more than four million English articles, we develop a pipeline to detect the quotations in the articles which are related to the experts and topics in the expert and topic corpora. As for Twitter, we conducted a particular procedure to query the experts' twitter accounts and retrieve their tweets that contain one of our controversial topics.

In the end, from these three sources, we have retrieved 35,720 arguments from experts related to the specific controversial topics until now.

Argument Stance Detection: Using the constructed SKG as the background knowledge, we propose an SKG-based model to predict the stances of the arguments from experts towards controversial topics (along with prediction's confidence scores).

The principle behind this model is that people's attitude to the topic has a decisive effect on their arguments about that topic, and the stance, which is indicated in the argument, is supposed to be consistent with their stance on the topic.

The SKG-based model can predict the stances of the expert's arguments in three situations: (1) the expert and topic of the argument are not identified in the SKG; (2) the stance of the expert towards the topic is identified in the SKG; (3) the expert's stance to the topic is not labeled in the SKG but the topic and the expert exist in the SKG.

In all the cases, for the input argument (by expert in controversial topic), the SKG-based model is supposed to provide the stance of the input argument as well as the corresponding confidence score.

Overall, from 35,720 mined quotations (regarded as arguments), the proposed SKG-based model was able to classify 26,524 arguments' stances into either Pro or Con. As for the rest, the model was not able to identify their stance because of the lack of background knowledge.

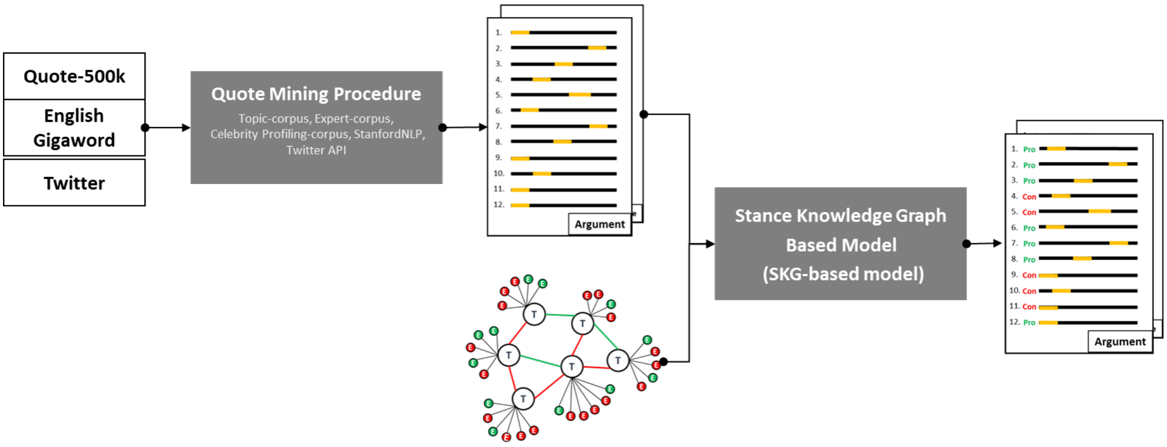
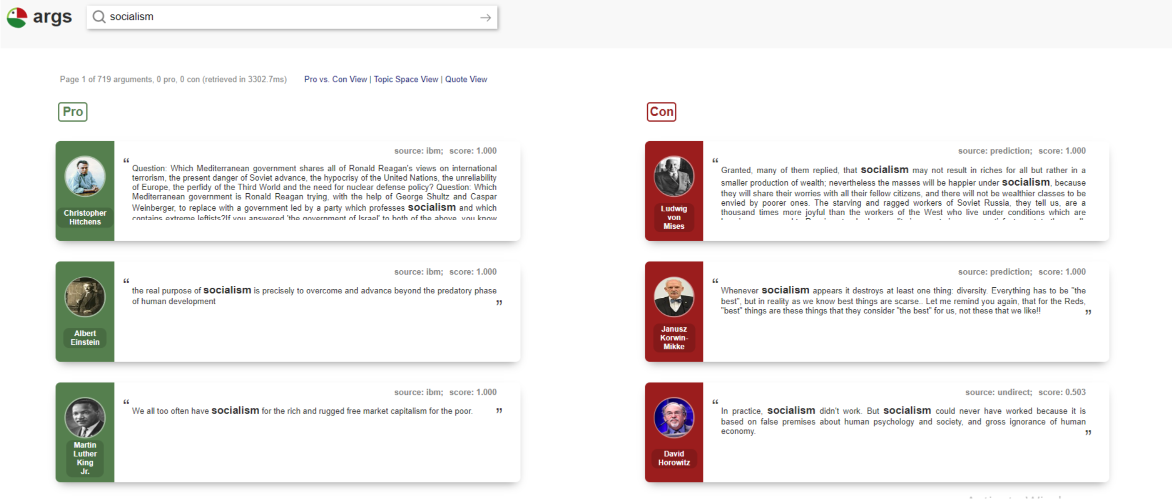

Quote-500k
English Gigaword
Twitter
Quote Mining Procedure
Topic-corpus, Expert-corpus, Celebrity Profiling-corpus, StanfordNLP, Twitter API
Argument
Stance Knowledge Graph Based Model (SKG-based model)
Pro
Con
Argument
T
E

Figure 3: Illustrates the ethos-evidence mining and stance classification

Integration on Argument Search Engines: The goal of this thesis is not only to automatically identify the stances of a set of mined arguments from the experts related to the controversial topics, but also to offer a solution for people to efficiently explore these arguments with different stances to get a trustworthy overview of the retrieved topic. In order to achieve this purpose, we integrated our work as a new module (*Quote View*) into Args.me.

Following the standard consideration and requirement of search engines, it is ideally required to evaluate the quality of arguments and rank them accordingly (in addition to the query relevance), which is essential for providing the users with high-quality arguments at the top of the search results. In this thesis, however, we adopted one criterion to measure and rank the retrieved arguments, namely, the confidence score, which is gained from the SKG-based model (described above). This score indicates how confident the model is regarding its stance prediction. Overall, all the arguments are ranked according to their confidence scores, where the argument with higher confidence score will be displayed on the top of results.

In the end, we evaluated and ranked 26,524 arguments from experts based on the confidence score and integrate them as one module into Args.me. This module is able to provide the users with ranked arguments from experts related to the queried controversial topic, and the retrieved arguments will be displayed separately according to their stances (either Pro or Con) as depicted in Fig 4.



Page 1 of 719 arguments, 0 pro, 0 con (retrieved in 3302.7ms)

source: ibm; score: 1.000
“Question: Which Mediterranean government shares all of Ronald Reagan's views on international terrorism, the present danger of Soviet advance, the hypocrisy of the United Nations, the unreliability of Europe, the perfidy of the Third World and the need for nuclear defense policy? Question: Which Mediterranean government is Ronald Reagan trying, with the help of George Shultz and Caspar Weinberger, to replace with a government led by a party which professes **socialism** and which contains extreme leftists?If you answered 'the government of Israel' to both of the above, you know

source: ibm; score: 1.000
“the real purpose of **socialism** is precisely to overcome and advance beyond the predatory phase of human development”

source: ibm; score: 1.000
“We all too often have **socialism** for the rich and rugged free market capitalism for the poor.”

source: prediction; score: 1.000
“Granted, many of them replied, that **socialism** may not result in riches for all but rather in a smaller production of wealth; nevertheless the masses will be happier under **socialism**, because they will share their worries with all their fellow citizens, and there will not be wealthier classes to be envied by poorer ones. The starving and ragged workers of Soviet Russia, they tell us, are a thousand times more joyful than the workers of the West who live under conditions which are

source: prediction; score: 1.000
“Whenever **socialism** appears it destroys at least one thing: diversity. Everything has to be "the best", but in reality as we know best things are scarse.. Let me remind you again, that for the Reds, "best" things are these things that they consider "the best" for us, not these that we like!!”

source: undirect; score: 0.503
“In practice, **socialism** didn't work. But **socialism** could never have worked because it is based on false premises about human psychology and society, and gross ignorance of human economy.”

Figure 4: Quote view in Args.me

1.3 Thesis Contribution

Altogether, the contributions of this thesis include: (1) constructing a new stance knowledge graph (SKG): the SKG encodes 82,501 experts' stances towards 170 controversial topics. It also encodes 604 topic correlation patterns with respect to the stance; (2) Mining 26,524 arguments from experts, identifying their stances using the SKG-based model, and integrating them into Args.me.

1.4 Thesis Structure

This thesis is structured as follows. In Section 2, we reviewed the existing studies in computational argumentation area. In Section 3, we introduced our construction of the background knowledge related to experts' stances towards controversial topics from Wikipedia to build and visualize the stance knowledge graph. In Section 4, we described the steps for mining arguments from experts, classifying their stances using the SKG-based model, and integrating them into Args.me. In Section 5, we concluded and discussed the future work.

2 Related Work

In this chapter, we reviewed existing studies in computational argumentation field, including ethos mining, argument scheme, argument mining, stance classification and quotation detection, which are relevant to our project.

2.1 Ethos Mining

Ethos is one of the persuasion strategies from the rhetoric theory proposed by Aristotle. It refers to the persona or perceived character of a speaker, whereas pathos refers to the audience's feelings triggered by the speaker and logos refers to arguments[18].

Ethos mining builds upon methods and techniques developed for sentiment analysis and opinion mining as well as argument mining and deep learning, which is summarized in [12]. Duthie et al. present a text analysis pipeline in [14] to identify expressions of ethos in political sphere and extract the information that expresses opinions during UK parliamentary debates. This pipeline is utilized to mine and analyze the statements in which speakers refer to other persons, named ethotic sentiment expressions (ESE). The stages in the pipeline include the ESE/Non-ESE stage, the +/-ESE stage and the network stage. In the first ESE/Non-ESE stage, there are three tasks: (1) parsing the plain text to the statements referring to agentive entity by implementing Named Entity Recognition (NER), Part-of-Speech (POS) tagging and a set of domain specific rules; (2) dealing with coreference by retrieving the person who utters the statement and the person who is described by the statement from the original text, and (3) discarding the statement which is not ethotic expressions but reported speech. The second stage uses a Support Vector Machine (SVM)-based sentiment classifier and an ethotic word lexicons to classify ESEs as pro or con. Next, in the network stage, the relationships between people, organizations and other entities will be displayed using the dataset from the previous stages. The pipeline that includes the ESE/Non-ESE stage and +/-ESE stage outperforms the baseline with an F1-score of 0.6 (0.4 higher than the baseline).

In 2018, Duthie et al., based on their previous work, introduce a deep modular recurrent neural network (DMRNN) in [13] for ESE/Non-ESE stage and +/-ESE stage and improve the classification performance with 0.84 F1-score for +/-ESE classification and 0.83 for ESE/Non-ESE classification.

2.2 Argument Scheme

An argument consists of several basic components, including a claim and premises. The claim is the central component which is supported by at least one premise, whereas the premise underpins the validity of the claim, as introduced in [36].

Walton et al. in [?] and [?] claim that argument from expert is one basic argumentation scheme with its own features. The argument from expert takes two premises which work together to support a claim (for example, the first premise: the expert is in the specific subject domain, the second premise: this expert thinks one proposition is true, conclusion: this proposition is true). It uses forwards reasoning to derive arguments from argumentation schemes and assumptions. Thus, for the argument from expert, its conclusion is a second-order variable ranging over propositions.

2.3 Argument Mining

Argument mining aims to automatically identify argument structure from unstructured natural language texts[24]. Basically, there are two methodologies in the argument mining field: context-independent approaches and context-dependent approaches.

Context-independent Approaches

Stab et al.(2014), Lippi et al.(2015) and Naderi et al.(2015), in [36], [24], [28], use the Support Vector Machine model (SVM) with various extracted features for argument mining. Stab et al.(2014) extract structural features based on the token statistics, including the location and the punctuation of the argument component as well as its covering sentence for identifying arguments. Lippi et al.(2015) extract features via the part-of-speech tagging (POS) for claim detection according to the common rhetorical structures observed in argumentative sentences based on the argument analysis literature. Their method reaches a precision of 0.098 and a recall of 0.587. Naderi et al.(2015) adopt word embedding models to extract features for argument mining in parliamentary discourse. [38] uses regression algorithms with extracted features composed by semantic density features, discourse and dialogue features and syntactic property features from the article for argument mining. Khatib et al.(2016) propose a distant supervision approach based on a binary statistical classifier with extracted features including token n-grams, discourse indicator, syntax and part of speech to robustly mine argumentative segments in discourses across domains in [3].

Neural network is also utilized for argument mining. Ajjour et al.(2017) in [1] present a Bi-LSTM based model using structural, syntactic, lexical and pragmatic features to segment argumentative discourse unit in cross-domain.

Context-dependent approaches

Levy et al.(2014) present an approach based on a cascade of classifiers (sentence component, boundaries component and ranking component) to detect context-dependent claims (CDC) related to the given topic from Wikipedia articles in [23]. The sentence component is developed using a logistic regression classifier (LR) to detect if the input text contains a CDC-sentence or not. The boundaries component is responsible for using Maximum Likelihood probabilistic model and another LR for detecting the exact CDC boundaries within the CDC-sentence from the previous step. The ranking component is based on another LR to score the CDC-candidates. Afterwards, Rinott et al. (2015) improve this approach and propose Context Dependent Evidence Detection (CDED) in [31] to detect the supportive evidence for the claims with LR from Wikipedia as well.

In [31], Hua et al.(2017) develop a system to mine the positive premises in the sentence level from relevant documents for a user-specified claim using logistic linear model. The model is utilized to find and locate the arguments in the cited document which can be used to back up the claim.

Neural network models are used to integrate context related information for argument mining. [37] from Stab et al.(2018) investigates cross-topic argument mining from heterogeneous sources, not limited to Wikipedia. In the paper, they created the manually annotated corpus with over 25,000 instances related to eight controversial topics from a variety of text types after removing boilderplate text and the sentences without verbs or with less than three tokens by adopting tokenization, sentence segemntation and part-of-speech tagging. Then, they propose the contextual BiLSTM model (biclstm) to integrate topic information as well as involve multi-task learning and transfer learning models setups to leverage additional data for the potential improvement of the performance. In the conducted cross-topic experiments, the biclstm model leads to better generalization to unknown topics, improving 5.9% in F1 in the two-label setup (argument or non-argument) and 4.6% in the three-label setup (pro-argument, con-argument or non-argument) over the the baseline. Their work showed the necessity of integrating topic information of arguments for argument mining and classification.

2.4 Stance Classification

Stance classification aims to automatically identify whether an input text it supportive, opposed or neutral towards a target entity as described in [33]. A variety of approaches have been investigated for the stance classification task.

Feature-based Machine Learning Models

Many studies on the Stance classification focus on the dialog level and involve the dialogical information in the proposed models. The study of [17] shows besides the training data size and model complexity, dialogic information can improve the performance of the stance classifier as well. Marilyn et al.(2012) in [44] model the authors' dialogic relations with a graph in which each node is a post and the edges indicate dialogic relations between posts (agreement or disagreement) and implement MaxCut method for the stance classification task. [30] proposes the Joint Viewpoint-Topic Model with User Interaction (JVTM-UI) to detect the viewpoint from threaded forum posts by considering viewpoint-based topic distribution (each viewpoint has its own distribution of topics), user identity (the viewpoint is from the corresponding user's viewpoint distribution) and user interactions (interaction expressions in the forum can have clue indicating agreement or disagreement). Scridhar et al.(2014) in [35] construct a stance classification model for posts from online debates and they use a probabilistic soft logic model to capture relational information about users and posts. These studies utilized background knowledge, such as the interaction between the authors of the input texts, for stance classification, but they focused on the dialog genre.

Neural Network based Models

Neural Network based approaches also play an important role in the stance classification task. [39] proposes a two-step model for the sentiment classification in the document level with the accuracy 0.651. In the first step, a long short-term memory (LSTM) is used to learn sentence representation and in the second step, a gated recurrent neural network (GRU) is used to encode semantics of sentences and their inherent relations in document representations which can be adopted as features to classify the sentiment label for each document. Besides that, the topic-involved stance classification was also studied. [4] from Augenstein et al.(2016) experiments with conditional LSTM encoding to model the tweet and the topic dependently and demonstrate that it outperforms tweet-topic independent encoding. [6] proposes a system working on the claim stance classification with respect to a given topic, including claim target identification with logistic regression classifier to detect the target of the given topic and claim, claim sentiment classification with lexicon-based sentiment analysis for the stance classification towards each of the targets, and contrast detection to check whether the targets are consistent or contrastive.

The attention mechanism introduced by Bahdanau et al.(2014) in [5], which weights the tokens of the input text based on their corresponding importance for the interpretation of the input text, is widely applied in the stance classification approaches. Zhou et al.(2017) in [49] propose to embed a novel attention mechanism in the bidirectional GRU-CNN neural network which can automatically capture semantic features in deciding the target-specific stance. Du et al.(2017) propose a BiLSTM model in [11] which incorporates with topic related information located by a novel attention mechanism. The introduced model on the Stance Detection dataset from English tweets outperforms the baselines. Dey et al.(2018) in [10] present a two-phase LSTM-based model with attention mechanism to label the user's stance shown in the twitter to the given topics. The first phase classifies the tweets into neutral or non-neutral class and the second phase classifies the non-neutral tweets into pro stance or con stance. The F-score reached 0.6884. In [45], Wei et al.(2018) propose a neural model for Twitter stance detection, named TGMN-CR (Target-Guided Memory Network with Conditional Representation), which uses two bidirectional gated recurrent unit networks and a soft attention mechanism to learn target-specific tweet representations and extracts stance-indicative rules via multiple interactions between topic and tweet words for the prediction of the stance distribution.

Based on the attention mechanism, Vaswani et al.(2017) propose the Transformer architecture in [42], which is implemented by Devlin et al.(2018). They introduce a language representation model named BERT (Bidirectional Encoder Representations from Transformers) with bidirectional attention mechanism in [9]. Mayfield et al.(2019) in [26] demonstrate that BERT model contextualized word embeddings consistently outperform other language representations. [29] introduces a consistency-aware neural network model based on BERT model to identify if the given perspective from the user is supporting or opposing the claim. Soleimani et al.(2020) in [34] introduce a three-step pipeline for evidence-based claim verification. The pipeline consists of document retrieval, sentence retrieval and claim verification. The document retrieval is to get the documents possibly containing relevant information to support or reject a claim. The sentence retrieval part uses BERT model to extract the top five evidence sentences. In the claim verification part, another BERT model is trained to verify if each of the five sentences is for or against the claim. The aggregation of five individual decisions will determine the final claim classification.

Instead of predicting when the adopted model's accuracy on validation set reaches its maximum, Wei et al.(2016) design a 'vote scheme' for prediction in [46] and it conducts ten parallel epochs with a convolution neural network based model for stance detection in tweets and outputs the label which appears most frequently as the predicted stance for the given tweet.

In conclusion, up to now, for stance classification task, researchers mainly focus on extracting various features or involving topic related information to improve the performance of the stance classifiers. Some studies utilized the dialogical features as the background knowledge for the stance classification only in the dialog domain. As for other domains, few researchers concentrate on using the background knowledge which indicate the stance of the author to the topic, the input text related to, for stance classification task.

2.5 Quotation Detection

Quotation detection aims to locate spans of quoted speech in text, described in [?]. Many studies focus on the rule-based approach for that. For example, Pouliquen et al.(2007) in [?] propose 3 generic quotation matching rules to automatically detect quotations in multilingual news. Krestel et al.(2008) introduce a system in [?] to locate quotations by analyzing verbs and syntactic patterns. It is worth noting that these approaches both have a high precision and a low recall.

In addition, [?] introduces a system combining one simple boundary classification with a greedy prediction strategy for quotation detection and its performance is competitive with the state of the art.

3 Stance Knowledge Graph

We are motivated to develop a system to automatically classify whether the input argument from the expert is supportive or opposed towards the corresponding controversial topic. Up to now, most of the studies focus on extracting features from the input text itself for the stance classification task, but using deeper background knowledge has been studied less.

Inspired by the Expert Stance Graph introduced in [40], we propose a large-scale knowledge graph, named Stance Knowledge Graph (SKG), which indicates experts' stances to controversial topics and topic correlation patterns in the stance perspective. This graph will be utilized as the background knowledge to identify the stance of the arguments from the expert regarding a controversial topic.

The introduced SKG is constructed following distant supervision techniques with the information extracted from Wikipedia related to the specific experts and topics. In the following, we describe it in detail.

3.1 Corpus Preparation

3.1.1 Seed Corpus

We built the seed corpus based on the Claim Stance dataset[40] containing 4603 Wikipedia categories and lists with annotated stances towards 132 two-sided controversial topics. After crawling the URLs of the Wikipedia page of each category or list, we noticed a fact that not only human entities were included in the Wikipedia categories and lists. At present, we focus on experts, so in order to filter out nonhuman entities, a humanity-check procedure via using Wikidata, an open knowledge base, is implemented. For each entity, we converted the entity's name to its unique id in Wikidata, then the id of the entity would be used to check if the type of entity is 'Human' or not. With this procedure, for each entity crawled from the Wikipedia categories and lists, we checked and only keep the human entities. In the end, the stances of 66,086 human entities towards 94 topics are collected as the seed corpus.

3.1.2 Topic Corpus

The 94 gained controversial topics from the seed corpus are the key topics. In order to expand topics, the Debate Topic Expansion dataset[7] was used to find related topics for each key topic. In this dataset, every related topic to the key topic was grouped into either the contrastive topic group or the consistent topic group and annotated as either a good expansion or a bad expansion for the key topic. Based on that, a structured dataset named Topic Net was created (a sample is shown in Table 1a), in which each key topic's contrastive and consistent topics labeled as the good expansion are included.

key topic	relation	related topic
atheism	consistent	irreligion
		secularism
		atheism
	contrastive	religion
		theology
		pantheism
		christianity

(a) Examples from Topic Net

topic	terms
authoritarianism	dictatorship
	absolutism
	caesarism
	despotism
	monocracy
	tyranny
	authoritarianism

(b) Examples from Topic Term

Table 1: Examples from the topic corpus illustrating the scheme

In addition to that, another dataset named Topic Term was built, which is used for the topics' terms that are different expressions or forms but referring to the same meaning or concept (see Table 1b). The topic term dataset is constructed from two parts, including the Claim Stance dataset[8] and WordNet[15]. In the Claim Stance dataset, each topic has one attribute called 'Category/List Page Title' and its value was treated as the terms of the given topic after a manual check to remove or correct non-term values, for example, the topic 'Abortion' has 'French_pro-choice activist','Irish_pro-choice-activists' and 'List of songs about abortion' in its attribute 'Category/List Page Title', but after the manual check, only 'pro-choice' from these three values will be kept as the term of the topic 'Abortion'. For the WordNet, the synonymous words of the given topic which are not identified as the topics in the Topic Net dataset and checked manually will be kept as the terms for this given topic.

Besides that, in consideration of the convenience for the further data process, the key topic itself is also one of its consistently related topics in the topic net dataset. As for the topic term dataset, the topic itself will be one of its terms as well.

The difference between topics and terms lies in that the topics describe or represent different concepts, but the terms of the same topic refer to the same concept in the different morphological expressions or forms, for example, the word 'Abortionist' is one of the terms of the topic 'Abortion' instead of one of its related topics.

3.1.3 Expert Corpus

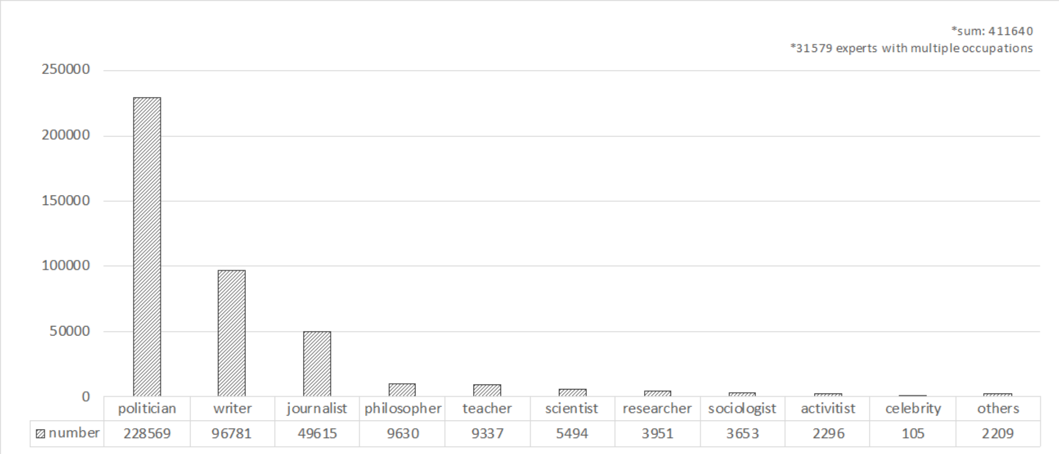
The expert corpus contains 446,166 experts which are collected from two data sources, the seed corpus and Wikidata. Table 2 lists the examples from the expert corpus and its scheme.

Expert	URL
John Dewey	/wiki/John_Dewey
Romano Prodi	/wiki/Romano_Prodi
Nigel Farage	/wiki/Nigel_Farage
Ruth Barnett	/wiki/Ruth_Barnett
Merle S. Goldberg	/wiki/Merle_S._Goldberg

Table 2: Examples from the expert corpus illustrating the scheme

In the seed corpus, there are 66,086 human entities, and these experts have the corresponding URLs from Wikipedia to identify them. As for Wikidata, in total, it encodes 4,391,071 human entities. However, due to the current demand and computation limitation, we take the human entities basically from 10 specific occupations into consideration at present. With this occupation restriction, there are 380,080 experts having English individual Wikipedia URLs left from Wikidata and these experts exclude the ones from the seed corpus.

The distribution of the occupations is indicated in Figure 5. We think people with these occupations (including politician, writer, researcher, journalist, scientist, activist, sociologist, celebrity, teacher, philosopher, etc.) are more likely to share their opinions about controversial topics to the public and have a higher impact on the attitude and belief of ordinary people.



	politician	writer	journalist	philosopher	teacher	scientist	researcher	sociologist	activitist	celebrity	others
▨ number	228569	96781	49615	9630	9337	5494	3951	3653	2296	105	2209

Figure 5: Occupation distribution

3.1.4 Expert Stance Evidence (ESE) Corpus

The Expert Stance Evidence (ESE) corpus was built by a pipeline for the extracted information in the sentence level from Wikipedia which involves the topics and the experts from the corresponding created corpus. These sentences, as the evidence of the experts' stances, are supposed to imply the relationship between the topic and the expert through the behaviors or moral beliefs of the expert shown in the sentence. This adopted pipeline contains 5 steps describing as follows:

(1) convert the Wikipedia database backup dumps, a complete copy of the information from the Wikipedia, into the structured dataset including the cleaned texts in the document level and their corresponding URLs by using WikiExtractor.

(2) utilize the Stanford CoreNLP[25] which provides linguistic analysis tools for the text data to implement the co-reference analysis and segment the documents into sentences. The reason to take the co-reference into consideration is that people tend to use the pronouns or other pro-forms to refer to the entities mentioned before, which is not problematic and ambiguous for understanding with the help of the context. However, after segmentation, the underlying binding between co-referential expressions will be removed. Thus, conducting co-reference analysis and replacing the pronoun with its antecedent are essential before the article segmentation so as to avoid discarding target sentences. Moreover, in order to improve the efficiency of this step, the topic corpus was used to filtered out irrelevant articles to reduce the workload for the co-reference analysis. The output of this step includes the sentences and the URLs of the Wikipedia articles in which each sentence can be found.

(3) implement the topic corpus to remove unrelated sentences that do not contain any term of the topics in the topic corpus. Afterward, the expert corpus will be used to discard the sentences without any expert in the expert corpus. In the end, for each filtered sentence, the expert, the key topic, the related topic, the terms appearing in that sentence, and the relation (either 'contrastive' or 'consistent') between the key topic and related topic will be annotated.

(4) conduct the URL-check procedure which can access the URL of the article where the given sentence exists and compare the expert's URL to each hyperlink in the Wikipedia webpage of the corresponding article. The purpose of the URL-check procedure is to guarantee that the expert mentioned in the sentence exists in the expert corpus instead of another person having the same name.

(5) annotate the stances of the sentences gained from the last step. Due to the different annotation methods, we divide the collected sentences into 3 sets, named IBM-ESE, Deduction-ESE, and Nonstance-ESE respectively. The IBM-ESE dataset is for the sentence where the mentioned expert's stance towards the topic in the sentence is directly labeled in the Expert Stance Graph, and the deduction-ESE dataset is for the sentence where the stance is not directly indicated in the Expert Stance Graph but can be deduced based on the assumption that the expert's stances should be the same to the consistently related topics and be the opposite to the contrastively related topics. The rest sentences whose stances can be neither directly acquired nor deduced comprise the nonstance-ESE dataset.

In the end, the IBM-ESE dataset contains 41,961 sentences with annotated stances from 7,072 experts towards 92 controversial topics and the deduction-ESE dataset has

4,287 sentences with annotated stances related to 446 experts and 56 topics, whereas the nonstance-ESE dataset includes 130,850 sentences without identified stance from 26,129 experts towards 170 topics.

We merged the IBM-ESE and the deduction-ESE to build a corpus called stance-ESE corpus and it will be used to train a stance classifier, whereas the nonstance-ESE will be used to scale the SKG. Examples from the ESE corpus are shown in Table 3.

key topic	**related topic**	**term**	**expert**	**sentence**	**stance**	**relation**
socialism	marxism	communists	karl marx	karl marx also proclaimed that the communists ...	pro	consist
apartheid	apartheid	anti-apartheid	bob marley	...bob marley support for ... anti apartheid movement	con	consist
judaism	judaism	jews	byron sherwin	a positive view of jesus is fairly represented among...	pro	consist

(a) Examples from Stance-ESE Corpus

key topic	**related topic**	**term**	**expert**	**sentence**	**relation**
evolution	evolution	evolution	john page	...john page hopps supported both evolution and spiritualism ...	consist
nazism	nazism	nazis	emmanuel	...emmanuel levinas wrote on jewish spirituality...	consist
anarchism	anarchism	anarchism	zhuangzi	in china philosophical anarchism... by taoist philosophers, zhuangzi and lao tzu ...	consist

(b) Examples from the Nonstance-ESE Corpus

Table 3: Examples from the ESE corpus illustrating the scheme

3.1.5 Corpus Evaluation

We conducted the distant supervision technique to build the ESE corpus which is used to train an ESE stance classifier, so the quality of generated ESE corpus will significantly influence the performance of the trained ESE stance classifier. Thus, it is important to evaluate the ESE corpus and analyze its quality.

3.1.5.1 Approach

As described before, the ESE Corpus contains the expert stance evidence (ESE) in the sentence level from Wikipedia. From the test dataset, we randomly selected 350 ESE sentences for the evaluation. Annotators classified the ESE sentences by using a browser-based interface which presents sentences, corresponding topics and experts, and a multiple-choice form (Pro, Con, others) to identify whether each sentence indicates the pro stance of the expert to the topic, the con stance of the expert, or no clear stance from the expert, respectively.

Two annotators mainly classify these 350 ESE sentences. Each annotator specified 200 sentences, and 50 overlapped sentences were classified by both annotators. As for the disagreement in the overlapped sentences, the third annotator would classify them and the final annotation results will be identified based on these three annotations following the majority rule.

3.1.5.2 Discussion

Inter-annotator agreement for our two annotators was 0.353 as measured by Cohen's kappa, which indicates fair agreement between our annotators, whereas the agreement between the annotators and labeled stances was moderate with Cohen Kappa value 0.447. However, nether of them exceeded 0.7 which is a threshold commonly used for assuming the results are reliable[**?**]. It is probably because that the number of the overlapped sentences is not big enough and it could not properly represent the real inter-annotator agreement.

From the 350 selected ESE sentences, there were 107 sentences annotated with the label 'others' after resolving the disagreements. As for the rest sentences, the result is shown in Table 4. The F1-macro value is 0.723, and the accuracy value is 0.929. If including the sentences which are classified in the 'others' group, the accuracy value will reduce to 0.643.

	Pcon	**Ppro**	**Rcon**	**Rpro**	**F1-macro**	**Accuracy**
ESE Corpus	.500	.959	.471	.964	.723	.929

Table 4: ESE corpus evaluation

3.2 ESE Stance Classifier

3.2.1 Features Extraction

To keep more semantic information of the input text, the non-English elements will be removed, including numbers and punctuation marks and so on. In addition to that, some stop words (such as 'Not') might play an important role in the stance classification, so these words will not be filtered out from the sentences.

Word Count Vector: the input text is notated as the term-document matrix[21], in which each row represents a sentence and each column represents a word in the vocabulary. The value of each cell shows the number of occurrences of a particular word (defined by the column) in a particular sentence (defined by the row).

TF-IDF Vector: Besides using simple frequency as the measure of association between words in the term-document matrix, tf-idf score which weights terms in the vector can be used to indicate the relative importance of a particular word in a particular sentence[21]. Regarding the different levels of input tokens, there are three different tf-idf vectors including word level vector (tfidf-word), n-gram level vector (tfidf-ngram) and character level vector (tfidf-char). The word level vector extracts only unigrams in the word level, but the n-gram level and character level extract unigrams and bigrams in the word level and character level respectively.

Word Embedding: For the input text, each word will be represented as a vector showing a point into multiple dimensional semantic where the similarity between words is indicated[21]. Two pretrained word embeddings were applied on the stance-ESE corpus, respectively, fastText (a context-free representation trained on Wikipedia 2017, UMBC web base corpus and statmt.org news dataset)[27] and BERT (a contextual representation trained on BooksCorpus and Wikipedia)[41].

3.2.2 Models Building

For the feature-based machine learning models, *Naive Bayes*[21], *Logistic Regression*, *Support Vector Machine (SVM)*[48] were experimented with extracted features (word count vector and tf-idf vector in different levels).

As for neural network models, pretrained word embeddings were adopted. *Convolutional Neural Network (CNN), Long short-term memory (LSTM), Gated Recurrent Units (GRU), Bidirectional LSTM (BiLSTM)* and *Recurrent Convolutional Neural Networks (RCNN)*[21] used the fastText for word embedding. The BERT word embedding was exclusively used for the *Bidirectional Encoder Representations from Transformers (BERT)*[9] model.

We experimented with these models (except BERT) using the corresponding extracted features under two input settings, namely, unmask setting and mask setting.

- **mask setting:** every term of the topic and the expert appearing in the sentence will be masked by the words 'topicWord' and 'expertWord' respectively.
- **unmask setting:** the term and the expert in the sentence will be revealed without masks.

As for the BERT-based model, we proposed another 3 different input settings which leverage various amounts of additional information related to the topic and expert. For each sentence S including the topic T and the expert E, the inputs under different settings are as follows.

- **setting A:** without involving the expert or the topic information, the input is formed into a text sequence [[CLS], S, ['SEP']].
- **setting B:** only with the topic related information, the input is formed into a text sequence [[CLS], T, ['SEP'], S, ['SEP']].
- **setting C:** with the expert and the topic related information, the input is formed into a text sequence [[CLS], T, E, ['SEP'], S, ['SEP']].

Due to the maximum length limit of the input, for each setting, the length of the sentence part in the input will be different and limited. In order to let the model gain the information more related to the topic and the expert, we select the part of the sentence which contains the information relevant to the topic and the expert as much as possible. To achieve that, the expert and each term of the topic in the sentence will be located. Then a window with the maximum length of the sentence part will be moved word by word from the first word to the end so as to collect all the sentence segments including the expert and at least one term of the topic. After getting all the target sentence segments, only the first one which has the highest occurrences of the expert and the topic terms will be selected as the input. If the expert and any term of the topic cannot exist together for every segment, the segment having the highest occurrence of terms of the topic will be chosen as the input for the model.

Moreover, it is worth noting that the stance-ESE corpus is extremely imbalanced in terms of the stance with the ratio 9:1. Because of the lack of the information about the ratio between pro and con stances to controversial topics in reality, we implemented oversample and undersample methods separately to balance the data with the ration 1:1 in the stance aspect.

- **oversample:** we trained only one classifier for the stance classification task.
- **downsample:** we randomly split the sentences with the pro stance in the training data into 9 sets. After merging each set with the con-stance sentences in the training data, we generated 9 balanced sub-training datasets.

With the downsample method, 9 sub-stance classifiers were trained based on the generated 9 balanced sub-training datasets correspondingly. Then one logistic regression layer was added to aggregate the final output.

3.2.3 Experiment Setup

The stance-ESE corpus with annotated stances was split into the training, the validation and the test datasets following the ratio 7:2:1 (see Table 5). Because some sentences include multiple topics or experts, in order to avoid the information leakage which can cause a negative influence on the classifier training, the sentences having more than one topic or expert will be all put into the training dataset to guarantee no shared sentence among the training, valid and test datasets.

	Train	**Validation**	**Test**	**Sum**
Pro	29,061	8,220	4,235	41,516
Con	3,321	931	480	4,732
Sum	32,382	9,151	4,715	46,248

Table 5: Distribution of the data for the stance classifier building

For machine learning models, we combine the training and validation datasets for the cross-validation to train and tune their hyperparameters, and the trained classifier with the lowest validation loss will be evaluated on the test dataset. For neural network models, the Adam optimizer[22] was used to tune the models and the model with the lowest validation loss will be adopted on the test dataset.

With the intention of accelerating the training process, the epoch number was limited to 5. Besides, for the BERT model, the maximum length of sequence input was limited to 128, and for others, every sentence was truncated at 70 words.

3.2.4 Discussion

As evaluation measures, we report the F1-macro value and each stance class's Precision value (Pcon, Ppro) and the Recall value (Rcon, Rpro).

The performances of the classifiers trained based on different machine learning models with different settings and features are represented in Table 6. With the same extracted feature, the F1-macro score in the experiment under the unmask setting is significantly higher than that under the mask setting.

Model	Setting	Features	Pcon	Ppro	Rcon	Rpro	F1-macro
Dummy(baseline)	unmask	word	.07	.92	.08	.92	.5
		tdidf-word	.07	.92	.07	.92	.49
		tdidf-ngram	.06	.92	.06	.92	.49
		tidif-char	.06	.92	.06	.92	.49
	mask	word	.09	.92	.09	.92	.5
		tdidf-word	.06	.92	.06	.92	.49
		tdidf-ngram	.09	.92	.1	.92	.51
		tidif-char	.07	.92	.06	.93	.49
Naive Bayes	unmask	word	.59	.95	.56	.96	.76
		tdidf-word	.33	.98	.82	.81	.67
		tdidf-ngram	.23	.97	.79	.7	.59
		tidif-char	.22	.97	.84	.67	.57
	mask	word	.35	.93	.37	.92	.64
		tdidf-word	.21	.95	.67	.71	.56
		tdidf-ngram	.2	.95	.7	.67	.55
		tidif-char	.17	.95	.72	.59	.5
Logistic Regression	unmask	word	.53	.97	.73	.93	.78
		tdidf-word	.52	.98	.82	.91	.79
		tdidf-ngram	.27	.95	.6	.82	.63
		tidif-char	.44	.97	.73	.9	.74
	mask	word	.29	.94	.48	.87	.63
		tdidf-word	.27	.94	.53	.84	.62
		tdidf-ngram	.21	.93	.49	.79	.58
		tidif-char	.25	.94	.51	.83	.61
SVM	unmask	word	.63	.95	.59	.96	.78
		tdidf-word	.9	.95	.57	.99	.84
		tdidf-ngram	.56	.93	.33	.97	.68
	mask	word	.31	.92	.34	.91	.62
		tdidf-word	.62	.92	.22	.98	.64
		tdidf-ngram	.52	.91	.15	.98	.59

Table 6: Results for each machine learning model on the test dataset

As for the classifiers trained based on neural network models (performances represented in Table 7), the classifiers with unmask setting performed better than that with mask setting as well. Besides that, the F1-macro scores for the trained neural network models are generally higher, compared with those for the trained machine learning classifiers.

Model	Setting	Pcon	Ppro	Rcon	Rpro	F1-macro
CNN	unmask	.65	.98	.87	.95	.85
	mask	.38	.95	.62	.88	.69
LSTM	unmask	.52	.98	.84	.91	.79
	mask	.33	.95	.61	.86	.67
GRU	unmask	.61	.98	.83	.94	.83
	mask	.33	.96	.69	.84	.67
BiLSTM	unmask	.46	.99	.92	.88	.77
	mask	.35	.96	.66	.86	.68
RCNN	unmask	.45	.99	.95	.87	.77
	mask	.34	.97	.74	.84	.68

Table 7: Results for traditional neural network models on the test dataset

Table 8 represents the performances of BERT-based trained classifiers. The F1-macro score in the experiments with the oversample method improves averagely 0.37 over that with the downsample method. Moreover, the BERT-based classifiers significantly outperform other trained neural network classifiers.

Model	Setting	Method	Pcon	Ppro	Rcon	Rpro	F1-macro
BERT	A	oversample	.83	.97	.76	.98	.89
		undersample (lr)	.58	.99	.91	.92	.83
	B	oversample	.77	.99	.94	.97	.92
		undersample (lr)	.75	.99	.94	.96	.9
	C	oversample	.93	.99	.96	.99	.97
		undersample (lr)	.83	.99	.96	.98	.94

Table 8: Results for BERT model on the test dataset

Thus, the BERT model with the oversample method was adopted for the stance classifier. In order to evaluate the generalization ability of the model to an unknown topic or an unknown expert, we conducted cross-topic and cross-expert experiments only for the oversample method. For the cross-topic experiments, we selected sentences related to 80% topics, which means 110 topics of 139 topics from the stance-ESE dataset, as training data and the rest 20% topics as the test data. In order to have a valid result, the topics were selected and grouped manually for each run to guarantee that the ratios between stances in the training data and test data are similar. The same requirement was applied on the cross-expert experiments and the experts were split into the training data and test data in the proportion of 8:2 as well.

We averaged the results of the 5 runs for cross-topic and cross-expert experiments respectively, which are listed in Table 9. We noticed the performances of the stance classifier in the cross-expert experiments are overall significantly better than that in the cross-topic experiments with the average F1-macro scores 0.72 and 0.60 respectively, showing that the model's generalization is better to the unknown expert than that to the unknown topics.

Model	Type	Setting	Pcon	Ppro	Rcon	Rpro	F1-macro
BERT	cross-topic	A	.24±.088	.92±.051	.41±.225	.83±.081	.58±.063
		B	.37±.134	.91±.050	.29±.161	.93±.046	.61±.077
		C	.34±.276	.91±.057	.40±.167	.82±.214	.59±.138
	cross-expert	A	.47±.103	.95±.005	.52±.064	.93±.048	.71±.033
		B	.51±.180	.95±.015	.56±.112	.93±.052	.72±.049
		C	.54±.174	.96±.005	.58±.075	.93±.058	.74±.052

Table 9: Results of the cross-topic and cross-expert with BERT Model

In conclusion, after comparing all the trained models with the different methods and settings, We adopted the BERT model under the setting C with the oversample method to build the stance classifier, which has the best performance in the within-domain test and the cross-expert test. For the within-domain test, its F1-macro value can reach 0.97, and for unknown topics or experts, the F1-macro values are around 0.59 and 0.74, respectively.

3.2.5 Result

We applied the trained ESE stance classifier on the nonstance-ESE dataset. For each sentence in the dataset, we firstly removed the non-English elements as well and selected the part of sentence which contains the topic-expert related information as much as possible due to the limit of the tokens for the model's input. Then the adopted classifier was used to predict the stance of the expert with respect to the topic based on the sentence and calculate the corresponding possibilities of both stances as the confidence scores.

3.3 SKG Construction

Compared with the Expert Stance Graph described in [40], we would like to construct the SKG indicating not only the stances from the experts towards the controversial topics but also the correlation between each pair of these topics. Thus, there are two levels in the SKG, named topic-topic level and topic-expert level respectively. For the former level, it encodes the correlation between each pair of topics, and people's stances should be the same to the positively correlated topics and be opposite to the negatively correlated topics. As for the topic-expert level, it encodes the experts' stances on the controversial topics.

3.3.1 Topic-Topic Level

There are three sequential steps to build the topic-topic level of the SKG using the Expert Stance Graph, the deduction-ESE, and the nonstance-ESE with predicted stances as the input. Figure 6 illustrates the pipeline briefly. The function of each step will be explained as follows.

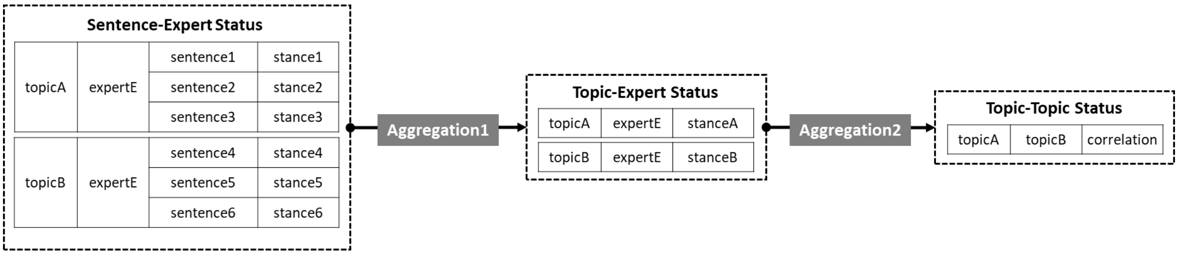


Figure 6: Graph Building illustration

Sentence-Expert status: In this step, for the sentences in nonstance-ESE dataset, their predicted stances were obtained with the corresponding confidence scores for both stances, represented as the ProScore and the ConScore respectively, by using the trained stance classifier described in the section 3.2. As for the deduction-ESE dataset and the Expert Stance Graph, they are already considered as the ground truth which directly indicates the experts' stances to the topics.

Topic-Expert status: The main task of this step is to identify the stance of the expert to the specific topic, but different aggregation methods are used for the data from the different datasets.

For the Expert Stance Graph, it is regarded as the ground truth and it directly indicates the experts' stances to the topics. Thus, for each topic-expert pair in this dataset, the confidence score for the labeled stance will be set to 1, whereas the confidence score for the opposite stance will be set to 0. In the deduction-ESE dataset, for each topic-expert pair, at least one evidence sentence was found and the expert's stance was annotated based on the principle that the expert's stances should be the same to the consistently related topics and be opposite to the contrastively related topics. Thus, this dataset will be also treated as the ground truth and the confidence score will be set to 1 for the annotated stance and 0 for the opposite stance as well.

As for the sentences with predicted stances in the nonstance-ESE corpus, The identification of the expert's stance to the topic is based on the multiple predicted stances of the corresponding ESEs as well as the their confidence scores (represented as ProScore and the ConScore). Specifically, we implemented the aggregation formula described in [2] showing as follows to predict the expert's stance to the given topic and calculate the possibility of the prediction as its confidence score:

$$Score(topic, expert) = \frac{\sum_{i=1}^{n}(ProScore_i - ConScore_i)}{n} \tag{1}$$

where n represents the number of the sentences related to the expert and the topic, the ProScore and ConScore represent the confidence scores indicating the possibility of the sentence with the pro stance and con stance separately. The method estimates the average of the difference between the ProScore and ConScore of the ESEs. If the score is bigger than 0, the expert's stance will be considered as a pro and if it is smaller than 0, the stance will be a con. Moreover, the score's range is from -1 to 1 and its absolute value indicates how possible the expert holds the predicted stance to the given topic. The bigger the absolute value is, the more confident the system is about its prediction.

Thus, in this step, we aggregated and converted the stances in the sentence level to the topic level by using the formula (1). The output indicates the experts' stances to the topics directly instead of the stances shown in the sentences and it will play an essential role for the topic-expert level in the SKG as well, which will be described later.

Topic-Topic status: This step focuses on concluding the correlation between topics, either positive or negative, with another aggregation method based on the output from the previous step.

We firstly collected all kinds of topic pairs with the possible stances and each pair can be represented as a pattern. For each topic pair, e.g. topicA and topicB, in consideration with the stances, there are 4 possible patterns, (topicA with pro, topicB with pro), (topicA with con, topicB with con), (topicA with pro, topicB with con), (topicA with con, topicB with pro), but there are only two possible correlations, positive or negative. The former two patterns represent the positive correlation and the latter two for the negative correlation. For this topic pair, we gathered all the experts who have identified stances towards these topics. Each expert' stances are indicated by the score(topicA, expert) and score(topicB, expert) where the positive value is translated as a pro and the negative value as a con following the formula (1). As for the topic pair's correlation and its confidence score shown from this expert, we conclude it by multiplying score(topicA, expert) and score(topicB, expert). The absolute value of the result represents how confident the system is about the predicted correlation and the positive value indicates the positive correlation, whereas the negative value represents the negative correlation. After considering all the experts having stances to these two topics with the formula (2), we aggregate and conclude the correlation between this topic pair.

$$Score(topicA, topicB) = \frac{\sum_{i=1}^{n} Score(topicA, expert_i) * Score(topicB, expert_i)}{n} \tag{2}$$

In this formula, the score(topic, expert) indicates the stance of the expert and its confidence score as described before, and n represents the number of the experts who have stances towards this pair of topics. The score ranges from -1 to 1. The negative score represents a negative correlation, whereas the positive score represents a positive correlation. As for the absolute value of the score, it indicates the possibility that the correlation between this topic pair exists. The higher the absolute value is, the more confident the system is about this aggregated correlation.

As for the occurrence of the aggregated correlation, a similar aggregation method was implemented. The number of the experts who have consistent stances on the topic pair will be set positive, whereas the number of the experts having contrastive stances will be set negative. After adding the positive number and negative number, the absolute value of the result will indicate how many times this correlation can be verified. It will be used as the criteria to filter out the uncommon patterns. Only when the number is bigger than the threshold 50, the pattern will be remained in the SKG, which means the pattern has been appeared more than 50 times among the experts.

In the end, after these three steps, we managed to gain the knowledge indicating the correlation between each pair of topics with the corresponding confidence score.

3.3.2 Topic-Expert Level

Based on the different methods used to identify the stance of the expert to the topic, there are three sets for the topic-expert level in the stance knowledge graph, including expert-ibm, expert-deduction and expert-prediction datasets. The basic statistical information was shown in Table 10.

	expert-ibm	**expert-deduction**	**expert-prediction**	**sum**
expert	66,086	446	26,141	82,503
topic	94	56	170	170
topic-expert pair	84,771	879	64,033	149,683

Table 10: Topic-Expert datasets

Expert-ibm dataset: The experts' stances with respect to the topics are directly identified in the Expert Stance Graph. Since the dataset was manually annotated and checked, it is regarded as the ground truth and its accuracy is the highest.

Expert-deduction dataset: The stance of the expert towards the given topic is based on the assumption that people's stances are the same for the consistent topics and the opposite for the contrastive topics. The stance from the expert to the given topic was deduced from the expert' stances to its related topics which were identified in the Expert Stance Graph.

The expert-deduction dataset is only composed of the experts with their stances to the topics when at least one sentence related to this expert-topic pair exists in the deduction-ESE corpus so as to avoid over deduction. Due to the lack of manual verification, the accuracy is not as high as it in the expert-ibm dataset.

Expert-prediction dataset: The experts' stances to the topics with the confidence scores were predicted and computed via using the trained stance classifier and the aggregation methods explained above in the topic-topic level building part.

In the end, the SKG encodes 604 high-frequent topic correlation patterns (positive or negative) in the stance perspective in the topic-topic level, as shown in Table 11a. Besides that, it also encodes 149,683 stances (Pro or Con) from 82,501 experts towards 170 topics in the topic-expert level, as shown in Table 11b.

topicA	**topicB**	**correlation**	**score**	**count**
communism	socialism	positive	0.997	2674
fascism	communism	negative	0.977	1831
lgbt	transgender	positive	0.999	927
judaism	faith	positive	0.935	279

(a) topic-topic level sample

topic	**expert**	**stance**	**score**
abortion	Ron Paul	con	1.0
	Mike Huckabee	con	1.0
	Lila Rose	con	1.0
	Warren Hern	pro	1.0

(b) topic-expert level sample

Table 11: Stance knowledge graph sample

3.4 SKG Visualization

The generated stance knowledge graph was visualized via using the arc diagram visualization technology with several implemented interaction methods (e.g. click, hover, etc.) so as to represent the correlations between each pair of topics as well as the experts with their stances, as depicted in Figure 7.

3.4.1 Topic-Topic Level Visualization

As can be seen in Figure 8, it demonstrates the topic-topic level of the SKG. Fig 8a represents its default status, whereas 8b shows the interaction status during the hovering the chosen topic. The visualization solution for the topic-topic level of the SKG will be interpreted as follows:

The yellow nodes illustrate the topics from top to bottom in alphabetical order and the nodes for expanded topics are marked with the black border.

Curves linking the nodes demonstrate the correlation between each pair of the topics. The curve colors, red and green specifically, mean the negative correlation and the positive correlation respectively. The solid curves are specific for the correlation between original topics and the dashed curves are for the topic pairs including at least one

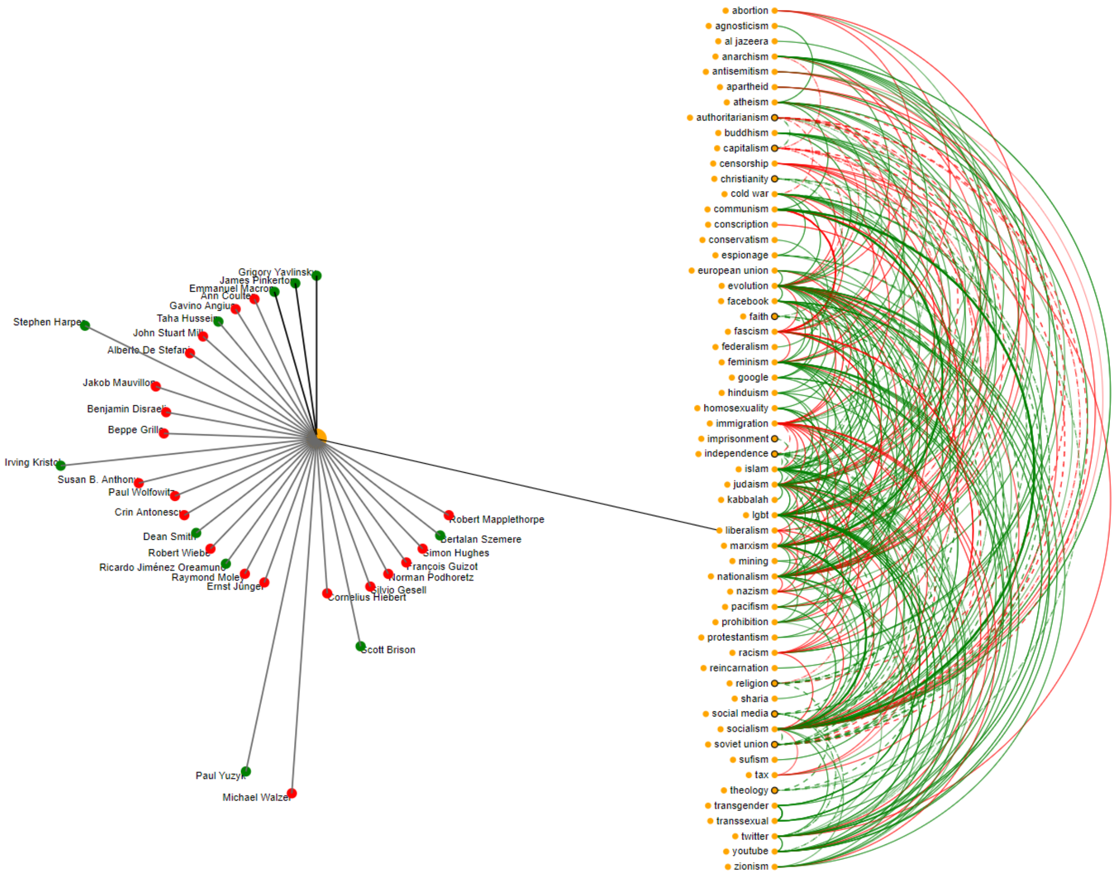
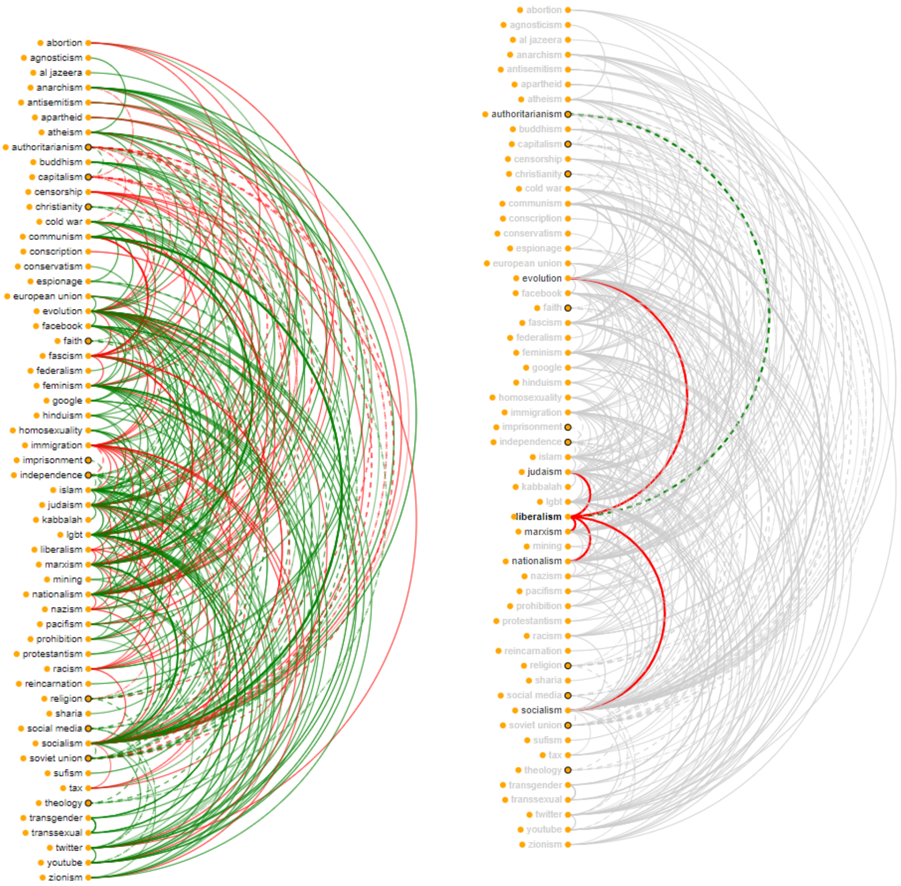


Figure 7: Stance Knowledge Graph visualization

extended topic. The opacity indicates the confidence score of the correlation. The lower the confidence score is, the more transparent the curve is. The width of the curve indicates the number of experts whose stances to the topics match the correlation.

As for the interaction, when hovering over one topic, the curves connecting the hovered topic and its correlated topics are highlighted as well as the corresponding topic nodes. As for other topic nodes, they will be dimmed, which is demonstrated in Figure 8b. As we can see, for the topic example 'liberalism', the SKG indicates there are 5 correlated topics including one extended topic, and it is negatively correlated with the topic 'authoritarianism' and positively correlated with the topics 'marxism', 'nationalism', 'socialism', 'judaism' and 'evolution'.



(a) default status (b) hover status

Figure 8: Topic-Topic Level visualization with interactions

3.4.2 Topic-Expert Level Visualization

As for each topic node in the topic-topic level visualization of the SKG, it contains multiple elements representing the experts with their identified or predicted stances to this topic, which indicates the topic-expert level of the stance knowledge graph.

As Figure 9 depicts, when one topic node is clicked, an additional diagram on the left of the topic-topic level visualization will appear. In this diagram, there is one node in the center representing the selected topic node and it is surrounded by the expert nodes with the experts' names.

The experts with their stances towards the selected topic come from three data sources, namely, the expert-ibm dataset, expert-deduction dataset and expert-prediction dataset, as described in section 3.3. The color of the link (black, dark gray, and light gray, respectively) between the expert node and the central topic node is used to distinguish the experts from three different data sources. Due to the limit of the space, we implemented stratified sampling to choose 30 experts. The distribution information can be roughly perceived in the central topic node from a pie chart which is accordingly formed based on the colors of the links to the experts. The expert node's color indicates the stance (red for negative stance and green for positive stance). The distance from the expert node to the central topic node indicates the corresponding confidence score. When the expert node is closer to the central node, it means it is more confident that the identified stance for this expert is correct and vice versa.

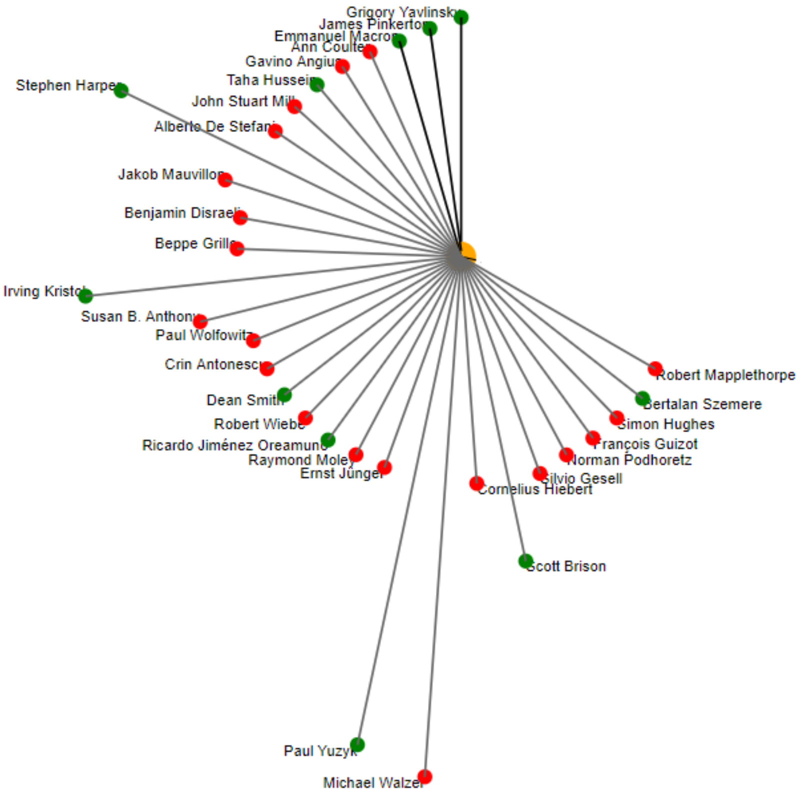


Figure 9: Topic-Expert Level visualization with interactions

3.5 SKG Evaluation

We constructed the SKG by aggregating and concluding the experts' stances showing in the sentence-level evidence. The quality of it has the decisive influence on its further application. Thus, it is necessary to evaluate the built SKG.

3.5.1 Approach

As described in 3.3, there are two levels in the built SKG. The topic-expert level encodes the experts' stances towards the topics, whereas the topic-topic level encodes the correlation between these topics. Thus, as for the expert's stance which can be predicted based on SKG, there are two possibilities, either the expert's stance is directly labeled in the topic-expert level or the stance can be identified based on the correlation between topics in the topic-topic level. Hence, the evaluation of the SKG can be divided into two parts accordingly.

For the first part where the stances are directly labeled in the SKG, we randomly selected 100 expert-topic pairs with the corresponding stances and these stances are all identified by the trained ESE stance classifier and the proposed aggregation methods. As for the stances which are identified based on the Expert Stance Graph, they are treated as the ground truth, so there is no need to evaluate them.

As for the second part where the correlations between topics in SKG are used to classify experts' stances, we only considered the expert's stance on the target topic which can be directly identified based on the the expert's stances on other topics in the SKG and the correlations between other topics and the target topic to avoid over deduction. We randomly selected 60 expert-topic pairs with the corresponding predicted stances.

Totally, there are 160 randomly selected topic-expert pairs. For each selected pair, we made a Google query to acquire documents related to the given topic and expert from heterogeneous sources, including news reports, blogs, forums, etc. The annotator was supposed to read these documents and specify whether the expert's stance on the topic is Pro, Con, or Unclear.

3.5.2 Discussion

It is worth noting that there are 149,683 possible topic-expert pairs from the topic-topic level and 1,048,575 pairs gained based on the topic-topic level. However, for the real application scenario, the former part functions the majority compared with the latter one. It is because every stance from the first part is identified based on the corresponding expert stance evidence in Wikipedia, but there is no single evidence from Wikipedia to prove the stance identified from the second part. Thus, for the topic-expert pair from the second part, the possibility is very high that there is no related information online, and accordingly, there are few arguments related to this topic-expert pair needing the stance classification. Hence, we analyze the topic-expert level and topic-topic level separately to evaluate the constructed SKG.

For the first part, among 100 topic-expert pairs, there are 79 pairs with annotated stances (either Pro or Con) and 21 pairs with unclear stances. As for the second part, among 60 topic-expert pairs, there are only 13 pairs with annotated stances (either Pro or Con) and 47 pairs with unclear stances. To evaluate the SKG, we only take the pairs with annotated pro or con stances into consideration.

Table 12 represents the evaluation results. The accuracy is 0.684 for the topic-expert level and 0.692 for the topic-topic level.

	Pcon	**Ppro**	**Rcon**	**Rpro**	**F1-macro**	**Accuracy**
Topic-Expert Level	.682	.684	.455	.848	.654	.684
Topic-Topic Level	.333	.800	.333	.800	.567	.692

Table 12: SKG evaluation

For the topic-expert level evaluation, we should notice that besides the stance gained by the prediction, there are still 57.2% experts' stances in the topic-expert level which are identified based on the Expert Stance Graph and they are treated as the ground truth. Taken the ground truth into consideration, the accuracy of the experts' stances in the topic-expert level can reach 0.864. If we consider the topic-expert pairs with unclear annotated stances (neither pro nor con) as false predictions, the accuracy of the experts' stances will reduce to 0.803.

As for the topic-topic level evaluation, due to the limited number of the topic-expert pairs with annotated stances, the validity of the evaluation cannot be guaranteed. However, in reality, the chance to use the topic-topic level of the SKG is rare. For example, among 26,524 mined arguments, only around 9% arguments' stances were identified based on the correlations from the topic-topic level.

Thus, for the mined arguments from experts, in consideration of the ratio between stances identified based on the topic-topic level and the topic-expert level, the accuracy of the experts' stances identified based on the SKG is 0.849 (removing topic-expert pairs with unclear annotated stances) and 0.793 (including topic-expert pairs with unclear annotated stances as false predictions).

4 Argument From Experts Mining

In this study, we focus on the mining arguments from experts. It is because, compared with ordinary people's arguments, experts' arguments are more convincing because of their deeper understanding in their professional fields as well as their ethos (credibility) which also functions as one persuasion strategy. At present, we hypothesis experts' quotations related to controversial topics are their arguments.

To automatically classify the stances of the arguments from the experts related to controversial topics, we propose an SKG-based model utilizing the built stance knowledge graph (SKG) described in section 3 as the background knowledge.

4.1 SKG-based Model

In order to classify the stances of the arguments from experts towards controversial topics as well as calculate the corresponding confidence score for the classifications, we propose an SKG-based model. It has two components. The first component is supposed to predict the given expert's stance to the given topic, whereas the second component could successively identify the stance of the arguments from that expert with respect to the specific topic.

4.1.1 Operating Principle

The principle behind this model is that people's stance towards the controversial topic should be consistent with their arguments' stances because people's stance on the topic has a decisive influence on the stance of their arguments and these arguments are just the medium representing their stance.

4.1.2 Classification Methods

In consideration of the expert, the topic of the argument, and the background knowledge encoded in the SKG, there are three different cases. The SKG-based model could implement different methods in different situations to classify the stances of the arguments from the experts.

Case 1: Either the expert or the topic does not exist in the SKG.

For the argument from an unknown expert or related to a new topic that does not exist in the SKG, our model is not able to predict the argument's stance. It is because neither the correlation between the given topic and any other topic in the SKG is identified, nor the given expert's stance to any other topic in the SKG is identified.

Case 2: The expert's stance to the given topic is identified in the SKG.

When the given expert's stance to the given topic is already identified in the SKG, the model will directly identify the stance of the argument with the confidence score as the same as the expert's stance to the topic with its confidence score showing in the SKG. The reason is that the stance of the expert's argument should be consistent with the expert's stance on its topic.

Case 3: The stance of the expert towards the topic is not identified in the SKG, but the expert and the topic exist in the SKG.

In order to identify the stance of the argument in this situation, the following procedures will be conducted step by step:

(1) Every topic which the given expert has the identified stance on in the SKG will be queried and represented as ($\mathrm{topic_{et}}$, $\mathrm{stance_{e}}$, $\mathrm{score_{e}}$) and the score is its corresponding confidence score indicating the confidence of the system about that the expert has the labeled stance towards the topic.

(2) Every topic which is correlated to the given topic in the SKG will be gathered and displayed as ($\mathrm{topic_{et}}$, target, $\mathrm{correlation_{t}}$, $\mathrm{score_{t}}$, $\mathrm{times_{t}}$), where the target represents the given topic, the score represents the confidence score indicating the possibility of the existence of the labeled correlation between the topic and the target, and the times indicates the occurrences of this correlation appearing among the experts in the SKG.

(3) With the related knowledge gained from the previous steps, for instance, ($\mathrm{topic_{et}}$, $\mathrm{stance_{e}}$, $\mathrm{score_{e}}$) and ($\mathrm{topic_{et}}$, target, $\mathrm{correlation_{t}}$, $\mathrm{score_{t}}$, $\mathrm{times_{t}}$) which indicate the expert's stance towards the $\mathrm{topic_{et}}$ and the correlation between the $\mathrm{topic_{et}}$ and the target respectively, we could identify the stance of the expert to the target topic and calculate its confidence score showing the possibility of the predicted stance. Specifically, the stances are supposed to be consistent towards positively correlated topics and be contrastive towards negatively correlated topics. Its corresponding confidence score is calculated based on the formula (3).

$$Score(expert, target) = score_e * score_t * (1 - e^{-times_t/100}) \tag{3}$$

where the $\mathrm{score_{e}}$ is the confidence score of the expert with the $\mathrm{stance_{e}}$ to the $\mathrm{topic_{et}}$, the $\mathrm{score_{t}}$ is the confidence score of the correlation between $\mathrm{topic_{et}}$ and target topic, and the $\mathrm{times_{t}}$ is the occurrence of the used correlation pattern in the SKG.

(4) Multiple topics might be correlated with the target topic, so different predicted stances with the corresponding confidence scores can be concluded. We split the output from the last step into two groups in terms of the predicted stance type (Pro, Con)

and select the one with highest confidence score in each group as the initial prediction for the Pro stance and the Con stance respectively, represented as (expert, target, pro, $\text{score}_{\text{pro}}$) and (expert, target, con, $\text{score}_{\text{con}}$). As for the situation where only one kind of initially predicted stance exists, either Pro or Con, the confidence score of the other stance that does not exist will be set to 0.

(5) Then, a calibration method is utilized for the confidence score correction in order to take the initial stance predictions into consideration for the final stance classification. The following formulas (formula (4) for pro and formula (5) for con) are conducted for the calibration, which can adjust the confidence score if there are conflicted initial predicted stances, but as shown in the formula, it cannot influence the confidence score when there is only one type of initial predicted stance.

$$Score_{pro}(expert, target) = \frac{score_{pro}^2}{score_{pro} + score_{con}} \tag{4}$$

$$Score_{con}(expert, target) = \frac{score_{con}^2}{score_{pro} + score_{con}} \tag{5}$$

In the end, the model will identify the stance of the input argument based on the calibrated confidence score. The stance with the higher confidence score will be identified as the expert's stance on the topic as well as the stance of all the arguments from this expert related to the corresponding topic.

4.2 Argument Retrieval

In this study, we hypothesis that the quotations and tweets including controversial topics from experts are very likely to be arguments and be persuasive for ordinary people due to the ethos. Thus, the experts' quotations are recognized as the arguments from the experts and crawling experts' quotations online is the strategy used for mining arguments from experts.

At present, we mined experts' quotations and tweets as their arguments from 'quote' datasets, such as Quotes-500k[32], news dataset, such as English Gigaword[16], and Twitter. The Quotes-500k created by Shivali et al.(2018) includes around 500 thousand crawled and cleaned quotations with speakers from four quote portals, including `goodreads.com`, `brainquote.com`, `famousquotesandauthors.com` and `curatedquotes.com`. David Graff et al.(2003) published English Gigaword which is a comprehensive archive of four distinct international newswire text data, containing more than 4 million English articles. Twitter is a platform where people could post messages (tweets). For different sources, different methods were adopted to mine the quotes from the experts related to the controversial topics in the corresponding corpus described in section 3.1 as arguments.

4.2.1 Arguments from Quotes-500k

For the Quotes-500k dataset, since the dataset was already cleaned and structured, a simple mining pipeline with two steps was adopted.

(1) use the topic corpus to filter out irrelevant quotes that do not have any term of the topics in the topic corpus. The key topic, the related topic, and the terms appearing in the quote will be annotated in the dataset.

(2) implement the expert corpus on the data gained from the previous step to get the quotes related to the specific topics from the experts in the expert corpus. The expert will be annotated as well.

4.2.2 Arguments from English Gigaword

The required quotations from the English Gigaword were mined and processed in a four-step pipeline.

(1) use the topic corpus to get the articles which contain at least one term of the topics in the topic corpus. The key topic, the related topic, the terms appearing in the article and the correlation between the key topic and related topic will be annotated for the filtered articles.

(2) adopt the expert corpus to process the topic-involved articles from the previous step and only keep the articles containing the experts from the expert corpus. The existing experts will be annotated as well.

(3) utilize the Stanford CoreNLP for quote detection to find the quotes as well as their speakers in the articles from the previous step. The quote will be taken if the detected speaker is consistent with the annotated expert and at least one term of the topics exist in the quote.

(4) implement part-of-speech tagging on the gathered quotes and remove the quotes without any verb or with less than three tokens.

4.2.3 Arguments from Twitter

For mining the tweets from the experts in the expert corpus related to the controversial topics, we conducted another 4-step pipeline to mine the target tweets.

(1) use the corpus introduced in [47], which links 71,706 verified twitter account ids with the corresponding Wikidata items and related personal information to match the twitter account ids with the experts in the expert corpus.

(2) find the twitter usernames which the experts are using at present with the account ids collected from the previous step.

(3) mine the tweets containing at least one term of the topics from the topic corpus for each twitter username we gathered. Due to the limit of the time and computation power, for each term, we maximally collect 10 tweets from one Twitter account.

(4) implement the part-of-speech tagging to remove the tweets without any verb or with less than three tokens.

After running the pipeline, we noticed that many experts sometimes use Twitter to promote or advertise activities or events with sharing the related links. For these tweets, they will have the words ’Twitter’, ’Facebook’, or ’YouTube’, but these words are not related to these tweets’ topics. Thus, the tweets containing the terms ’Twitter’, ’Facebook’, or ’YouTube’ will be dismissed.

In total, as can be seen in Table 13, we mined 35,720 quotes and tweets as the experts’ arguments towards controversial topics.

	Expert	**Topic**	**Argument**
Quotes **(Quotes-500k & English Gigaword)**	7703	131	13,614
Tweets (Twitter)	830	143	22,106

Table 13: Mined arguments statistics

4.3 Result

In the end, among 35,720 mined arguments from the experts related to the topics existing in the SKG domain, 26,524 arguments’ stances are classified into either Pro or Con stance with the corresponding confidence scores by using the SKG-based model. As for the rest, the model is not able to identify their stance because of the lack of valid relevant background knowledge.

4.4 Evaluation

These experts’ arguments with predicted stances are integrated as one module in Args.me. It is essential to evaluate them to assess the validity of this module.

4.4.1 Approach

As described above, there are 26,524 mined arguments from experts with predicted stances by using the SKG-based model. Among these arguments, we randomly selected 100 arguments mined from the quote-related datasets and 150 arguments from Twitter for evaluation. For each argument, the annotator should classify whether the argument is supporting the topic, opposing the topic or unrelated to the topic and annotated it with the label Pro, Con and Unrelated, respectively.

4.4.2 Discussion

For the randomly selected 100 arguments from the quote-related datasets, 66 arguments were annotated with either Pro or Con, and only 46 arguments from the 150 random arguments from Twitter were grouped into either pro or con.

According to our analysis of the rest arguments, we grouped these annotated unrelated arguments into three situations. (1) some arguments contain the terms of the topics but they unrelated to the target topics. (2) some experts in Twitter advertised events about the target topics and did not present their stances. (3) Some experts quoted other people's words about the target topics on Twitter instead of expressing their own arguments and showing their stances.

To solve this problem and improve its performance, we think enriching and improving our argument mining strategy could contribute significantly. At present, we implemented a basic string matching algorithm to mine the arguments related to the target topics without using further advanced methods, such as topic analysis techniques and machine reading comprehension models.

As for the arguments annotated with either Pro or Con, the evaluation result is shown in Table 14. The performance of the stance classification on arguments from Twitter is better than that from the Quote related datasets, and the accuracy are all higher than 0.77. If we group the unrelated arguments as the false stance prediction, the accuracy will significantly reduce.

	Pcon	**Ppro**	**Rcon**	**Rpro**	**Macro-F1**	**Accuracy**
Quote-related datasets	.783	.767	.643	.868	.760	.773
Twitter	.632	.926	.857	.781	.787	.804

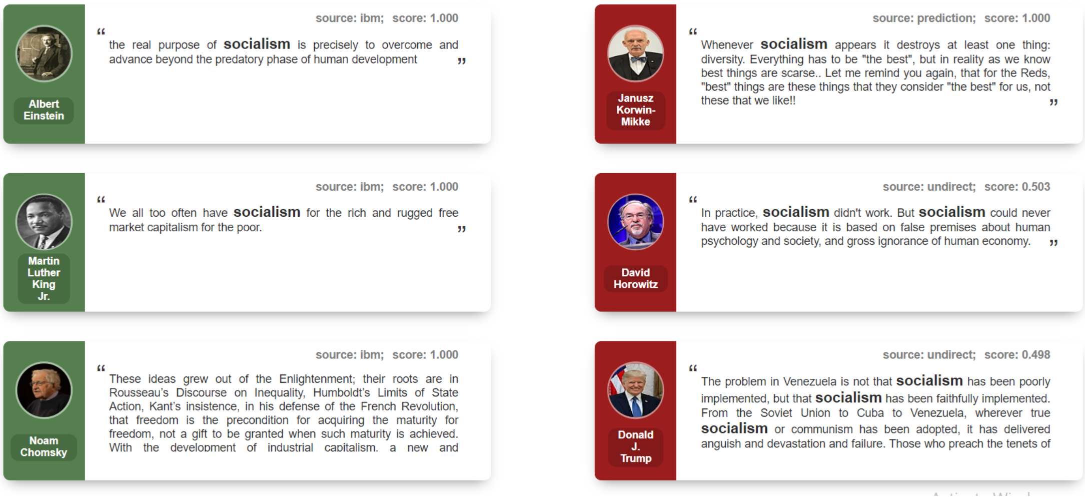
Table 14: SKG-based model evaluation

4.5 Application

Besides automatically identifying the stances of the arguments from the experts related to the controversial topics, we would also like to offer a solution for people to explore these arguments efficiently so as to get a trustworthy overview about it. Thus, we integrate our system into Args.me (an argument search engine introduced in [43]) working as one module, named quote view module.

4.5.1 Quote View Module

The quote view module could provide the arguments from experts related to the controversial topic the users are querying in Args.me. The interface of the module is illustrated in Figure 10.



Albert Einstein

source: ibm; score: 1.000

“the real purpose of **socialism** is precisely to overcome and advance beyond the predatory phase of human development”

Martin Luther King Jr.

source: ibm; score: 1.000

“We all too often have **socialism** for the rich and rugged free market capitalism for the poor.”

Noam Chomsky

source: ibm; score: 1.000

“These ideas grew out of the Enlightenment; their roots are in Rousseau’s Discourse on Inequality, Humboldt’s Limits of State Action, Kant’s insistence, in his defense of the French Revolution, that freedom is the precondition for acquiring the maturity for freedom, not a gift to be granted when such maturity is achieved. With the development of industrial capitalism, a new and

Janusz Korwin-Mikke

source: prediction; score: 1.000

“Whenever **socialism** appears it destroys at least one thing: diversity. Everything has to be "the best", but in reality as we know best things are scarse.. Let me remind you again, that for the Reds, "best" things are these things that they consider "the best" for us, not these that we like!!”

David Horowitz

source: undirect; score: 0.503

“In practice, **socialism** didn't work. But **socialism** could never have worked because it is based on false premises about human psychology and society, and gross ignorance of human economy.”

Donald J. Trump

source: undirect; score: 0.498

“The problem in Venezuela is not that **socialism** has been poorly implemented, but that **socialism** has been faithfully implemented. From the Soviet Union to Cuba to Venezuela, wherever true **socialism** or communism has been adopted, it has delivered anguish and devastation and failure. Those who preach the tenets of

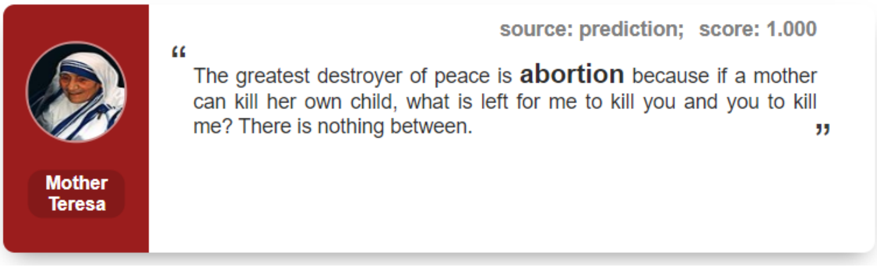
Figure 10: Quote View

4.5.1.1 Interface Layout

As shown in Figure 10, we adopted the Two Column Layout design. The retrieved arguments with respect to the given topic will be separately displayed based on their predicted stances, either Pro or Con. In other words, all the arguments will be split either in the Pro stance group or in the Con stance group. For each group, the arguments will be ranked following a ranking method which will be introduced later. The purpose of it is to help users to find the arguments they need as well as to get an overview of the topic efficiently.

4.5.1.2 Interface Design

For the interface design, we implemented the card-based UI (User Interface) design. Each card, as depicted in Figure 11, represents one expert and it provides the arguments from this expert as well as the basic information about the expert.



source: prediction; score: 1.000

“The greatest destroyer of peace is **abortion** because if a mother can kill her own child, what is left for me to kill you and you to kill me? There is nothing between.”

Figure 11: Quote Card

The card can be divided into two parts. The part on the left encodes the expert's information, including the name and the expert's photo queried from Wikipedia. When there is no available expert's photo in Wikipedia, the icon of Args.me will be used to represent this expert. The color of this part indicates the stance of the expert's arguments, red for Con stance and green for pro stance, which is consistent with other modules in Args.me.

The part on the right side is for the arguments related information. The terms related to the topic appearing in the argument text will be highlighted to improve the user experience. Besides the argument itself, it will also provide the confidence score of the argument to represent how confident the system is about the predicted stance, as well as the source showing how the stance is identified and concluded. For example, the source can be ibm, deduction, or prediction, which means the identified stance is based on the expert-ibm dataset, expert-deduction dataset and expert-prediction dataset, respectively, in the topic-expert level of the SKG. When the source is indirect, that indicates the stance is identified based on the correlation between topics encoded in the topic-topic level of the SKG.

4.5.2 Ranking Method

For the queried controversial topic, there might be many retrieved arguments from experts. To provide users with high-quality arguments at the top of the search results, evaluating and ranking the retrieved arguments are necessary and important.

In this thesis, at present, we adopted one criterion, the confidence score, to evaluate and rank the retrieved arguments. It is the output of the SKG-based model, as described above, besides the predicted stance of the input argument. It indicates how confident the system is about its prediction of the input argument' stance.

All the retrieved arguments will be ranked according to their confidence scores. The argument with higher confidence score will be given higher priority and displayed on the earlier. The arguments with the same confidence score will be ranked randomly. As for the arguments from the same experts towards the same topic, they will be grouped together in consideration of user experience. We noticed that the difference between retrieving arguments from experts and those from ordinary people is that experts' credibility could play an important role. Grouping the arguments from the same expert together is good for users to efficiently find the experts they trust and gain their arguments accordingly.

In conclusion, with the introduced ranking method, we evaluate and rank 26,524 experts' arguments. The finally ranking result will be indicated in the quote view module interface.

5 Conclusion and Future Work

In conclusion, we have presented a system to automatically identify the stances of the experts' arguments gained from their quotations based on the built stance background knowledge and integrate it into Args.me to provide a solution for people to efficiently explore these arguments from the respected experts and get a trustworthy overview about the topic they are interested in. Specifically, there are two contributions, (1) the stance knowledge graph, and (2) the arguments from experts mining.

Stance Knowledge Graph (SKG): We enriched the Expert Stance Graph introduced in [40] and built the SKG encoding not only experts' stances to controversial topics but also topic correlation patterns in terms of the stance.

In order to gain this knowledge, firstly we generated the expert corpus and topic corpus to identify the domains of the SKG in the expert and topic dimensions. The expert corpus was built based on the Expert Stance Graph after discarding non-human entities and including more experts from Wikidata with specific occupations, whereas the topic corpus was built according to the Debate Topic Expansion dataset[7] and the WordNet[15] besides the Expert Stance Graph. Based on that, we mined the information from Wikipedia in the sentence level which contains the entity from expert and topic corpora to form the expert stance evidence (ESE) corpus. Following distant supervision techniques, the ESE corpus can be divided into two parts, the stance-ESE corpus for the sentences having annotated stances based on the Expert Stance Graph and the nonstance-ESE corpus for the sentences having no labeled stance.

Secondly, we used the stance-ESE corpus to train an ESE stance classifier based on a BERT model (Bidirectional Encoder Representations from Transformers) to learn to extract the relation between the topic and expert from the given sentence. This trained classifier outperforms the baselines with 0.97 f1-macro for within the topic-expert test as well as 0.59 for the cross-topic domain and 0.74 for the cross-expert domain. For each sentence in the nonstance-ESE corpus, the trained stance classifier was used to predict the stance of the expert to the topic showing in that sentence.

After that, for each topic-expert pair, we aggregated and concluded the expert's stance towards the topic and its corresponding confidence score based on the predicted stances of the related ESEs. The predicted stances and annotated stances of the experts with the corresponding confidence scores form the topic-expert level of the SKG. Furthermore, we calculated and concluded the correlation between each pair of topics and its confidence score, which consists of the topic-topic level of the SKG.

In the end, we implemented the arc diagram visualization technology as well as necessary interaction techniques to visualize the generated SKG which encodes 82,501 experts' stances to 170 controversial topics and 604 high-frequent topic correlation patterns with respect to the stance.

In the future, the SKG can be scaled in the expert and topic dimensions so as to involve more experts and controversial topics, which can significantly improve the performance of the SKG-based model. The stance classifier for ESE could be improved by training with leveraging additional information or with a larger training dataset so as to improve the validity and reliability of the SKG.

Argument from Experts Mining: we proposed an SKG-based model to predict the stance of the expert's argument related to the controversial topic. The operating principle of this model is that the expert's stance to a controversial topic has a decisive effect on his or her related argument's stance and it should be consistent with the expert's stance to the topic, because the arguments are just the medium indicating the expert's stance. Thus, accordingly, the SKG-based model has two components. The first one is to predict the given expert's stance to the given topic with its confidence score based on the SKG, whereas the second component could successively identify the stance of the argument from that expert with respect to that specific topic.

The proposed model was applied to 35,720 experts' arguments retrieved from Quotes-500k[32], English Gigaword[16] and Twitter. Among them, 26,524 arguments' stances are classified into either Pro or Con stance with the corresponding confidence score.

Then, we integrated our work into Args.me and provide the quote view for users to query experts' arguments with stances related to topic they queried and efficiently get a trustful overview of it. We adopted the two column layout and implemented the card-based interface design for the quote view. The retrieved arguments in this view will be provided and displayed separately based on their stance.

Following the standard consideration and requirement of search engines, we adopted a ranking method to evaluate and rank all retrieved experts' arguments. At present, the confidence score which is gained from the SKG-based model is used to measure and rank arguments. It indicates how confident the system is about its stance prediction. The argument with higher confidence score will be given higher priority and displayed on the top. As for the arguments with the same confidence score, they will be ranked randomly and the arguments from the same experts towards the same topic will be grouped together in consideration of user experience.

In the future, for the argument mining, more sources and formats of high-ethos involved information could be taken into consideration, such as the experts' speeches, articles, comments, and so on, in order that the system could provide more experts' arguments about the controversial topics. Moreover, besides working on the quantity of the mined arguments, improving the quality of the argument is also one research direction to maintain a required accuracy of retrieved arguments and guarantee that the mined arguments are highly relevant to the topics.

As for the ranking method, it could be optimized by offering various ranking options to meet users' requirement, for example, ranking the arguments based on the expert's speciality, reputation, date, etc. Accordingly, the necessary interaction techniques could be implemented to improve the user experience.

References

[1] Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, 2017.

[2] Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner. Automatic detection of point of view differences in Wikipedia. In *Proceedings of COLING 2012*, pages 33–50, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[3] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1395–1404, 2016.

[4] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain, April 2017. Association for Computational Linguistics.

[7] Roy Bar-Haim, Dalia Krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 977–990, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Edward Loper Bird, Steven and Ewan Klein. *Natural Language Processing with Python*, 2009. O'Reilly Media Inc.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding, 2018. cite arxiv:1810.04805Comment: 13 pages.

[10] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, pages 529–536. Springer, 2018.

[11] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence, 2017.

[12] Rory Duthie. Recognising ethos in natural language. *3rd Summer School on Argumentation: Computational and Linguistic Perspectives*, page 17, 2018.

[13] Rory Duthie and Katarzyna Budzynska. A deep modular rnn approach for ethos mining. In *IJCAI*, pages 4041–4047, 2018.

[14] Rory Duthie, Katarzyna Budzynska, and Chris Reed. *Mining Ethos in Political Debate*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 299–310. IOS Press, Netherlands, 2016. This research was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National Science Centre under grant 2015/18/M/HS1/00620.

[15] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

[16] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.

[17] Kazi Saidul Hasan and Vincent Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, 2013.

[18] Colin Higgins and Robyn Walker. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting Forum*, volume 36, pages 194–208. Taylor & Francis, 2012.

[19] Xinyu Hua and Lu Wang. Understanding and detecting supporting arguments of diverse types. *arXiv preprint arXiv:1705.00045*, 2017.

[20] David J Jackson and Thomas IA Darrow. The influence of celebrity endorsements on young adults' political opinions. *Harvard International Journal of Press/Politics*, 10(3):80–98, 2005.

[21] Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[23] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[24] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[25] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[26] Elijah Mayfield and Alan W Black. Stance classification, outcome prediction, and impact assessment: Nlp tasks for studying group decision-making. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 65–77, 2019.

[27] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[28] Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *Principles and practice of multi-agent systems*, pages 16–25. Springer, 2015.

[29] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Stancy: Stance classification based on consistency cues. *arXiv preprint arXiv:1910.06048*, 2019.

[30] Minghui Qiu and Jing Jiang. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[31] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[32] Shweta Garg Shivali Goel, Rishi Madhok. Proposing contextually relevant quotes for images. In *40th European Conference on Information Retreival*, 2018.

[33] Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 159–169, 2016.

[34] Amir Soleimani, Christof Monz, and Marcel Worring. Bert for evidence retrieval and claim verification. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 359–366, Cham, 2020. Springer International Publishing.

[35] Dhanya Sridhar, Lise Getoor, and Marilyn Walker. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117, 2014.

[36] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, 2014.

[37] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[38] Reid Swanson, Brian Ecker, and Marilyn Walker. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226, 2015.

[39] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[40] Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. Expert stance graphs for computational argumentation. pages 119–123, 01 2016.

[41] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *ArgMining@EMNLP*, 2017.

[44] Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT ’12, page 592–596, USA, 2012. Association for Computational Linguistics.

[45] Penghui Wei, Wenji Mao, and Daniel Zeng. A target-guided neural memory model for stance detection in twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[46] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California, June 2016. Association for Computational Linguistics.

[47] Matti Wiegmann, Benno Stein, and Martin Potthast. Celebrity profiling. 08 2019.

[48] Hiroyasu Yamada and Yuji Matsumoto. Statistical dependency analysis with support vector machines. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 195–206, Nancy, France, April 2003.

[49] Yiwei Zhou, Alexandra I Cristea, and Lei Shi. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, pages 18–32. Springer, 2017.