

Investigating Rumor News Using Agreement-Aware Search

Jingbo Shang¹, Jiaming Shen¹, Tianhang Sun¹, Xingbang Liu¹,
Anja Gruenheid², Flip Korn³, Ádám D. Lelkes³, Cong Yu³, Jiawei Han¹

¹Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

²Google Inc., Madison, WI, USA

³Google Research, New York, NY, USA

¹{shang7, js2, ts7, xl14, hanj}@illinois.edu

^{2,3}{anjag, flip, lelkes, congyu}@google.com

ABSTRACT

Recent years have witnessed a widespread increase of rumor news generated by humans and machines. Therefore, tools for investigating rumor news have become an urgent necessity. One useful function of such tools is to see ways a specific topic or event is represented by presenting different points of view from multiple sources. In this paper, we propose Maester, a novel agreement-aware search framework for investigating rumor news. Given an *investigative question*, Maester will retrieve *related* articles to that question, assign and display top articles from *agree*, *disagree*, and *discuss* categories to users. Splitting the results into these three categories provides the user a holistic view towards the investigative question. We build Maester based on the following two key observations: (1) relatedness can commonly be determined by keywords and entities occurring in both questions and articles, and (2) the level of agreement between the investigative question and the related news article can often be decided by a few key sentences. Accordingly, we use gradient boosting tree models with keyword/entity matching features for relatedness detection, and leverage recurrent neural network to infer the level of agreement. Our experiments on the Fake News Challenge (FNC) dataset demonstrate up to an order of magnitude improvement of Maester over the original FNC winning solution, for agreement-aware search.

CCS CONCEPTS

• Information systems → Information retrieval; Specialized information retrieval;

KEYWORDS

Rumor News; Relatedness Classification; Agreement Detection.

ACM Reference Format:

Jingbo Shang, Jiaming Shen, Tianhang Sun, Xingbang Liu, Anja Gruenheid, Flip Korn, Adam D. Lelkes, Cong Yu, Jiawei Han. 2018. Investigating Rumor News Using Agreement-Aware Search. In The 27th ACM Int'l Conference on Information and Knowledge Management (CIKM'18), Oct. 22–26, 2018, Torino, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3269206.3272020>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '18, October 22–26, 2018, Torino, Italy

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-6014-2/18/10...\$15.00

<https://doi.org/10.1145/3269206.3272020>

The screenshot shows a search interface for 'Did Robert Plant turn down a contract to tour with Led Zeppelin?'. The results are categorized into three main sections: 'Agreeing articles', 'Disagreeing articles', and 'Discussing articles'. Each section contains a list of news articles with their titles, URLs, and brief summaries.

Agreeing articles	Disagreeing articles	Discussing articles
<p>Robert Plant turns down \$800 million for Zeppelin reunion www.cnn.com/2014/11/10/showbiz/according-to-a-report-in-the-uk-daily-mirror-the-zeppelin-lead-singer-turned-down-a-500-million-800-million-contract-for-a-led-zeppelin-reunion...</p>	<p>No, Robert Plant did not rip up an \$800 million offer to reunite Led Zeppelin https://consequenceofsound.net/.../no-robert-plant-did-not-rip-up-an-800-million-offer-to-reunite-led-zeppelin The Mirror recently reported that Robert Plant had turned down an \$800 million offer ... Turns out none of that actually happened ...</p>	<p>Led Zeppelin Not Dumb Enough to Turn Down \$800 Million www.hollywoodreporter.com/news/robert-plant-turned-down-800-747772 In a story that will only add to the legend of Led Zeppelin, the band's lead singer, Robert Plant, reportedly turned down the chance for the surviving members ...</p>
<p>Robert Plant Reportedly Turns Down \$800 Million for Led Zeppelin Reunion www.online.com/.../robert-plant-reportedly-turns-down-800-million-for-led-zeppelin ... Robert Plant has reportedly turned down more than \$800 million to reunite with Led Zeppelin on a worldwide tour.</p>	<p>Confused by Led Zeppelin rumours https://www.virgin.com/richard-branson/...confused-by-a-story-doing-the-rounds-about-us-apparently-offering-led-zeppelin-500-million-to-reform-and-carry-out-a-tour-as-much-as-i-love-the-band-there-is-absolutely-no-truth-to-the-story</p>	<p>Did Robert Plant Turn Down \$14 Million for Led Zeppelin Desert Trip ultimateclassicrock.com/led-zeppelin-desert-trip-reunion/ ... Robert Plant has gone on record repeatedly in recent years ...</p>
<p>Did Robert Plant Really Turn Down \$800 Million For A Led Zeppelin wxr.radio.com/.../did-robert-plant-really-turn-down-800-million-for-a-led-zeppelin ... a UK tabloid reported that Robert Plant turned down a 500 million pound ...</p>	<p>Did Robert Plant Tear Up \$800 Million LED ZEPPELIN Reunion Contract www.metalinjection.net/.../robert-plant-took-up-800-million-led-zeppelin-reunion-con According to UK tabloid The Mirror, all three living members of the Led Zeppelin ...</p>	<p>Led Zeppelin Reunion 2017: One Thing In The Way http://crazy4rock.com/.../i-love-rock-n-roll ... fans hear rumors of a Led Zeppelin reunion 2017, Robert Plant keeps standing in the way.</p>
<p>Update: Robert Plant's Publicist Denies Singer Turned Up \$800 Million ultimateclassicrock.com/robert-plant-led-zeppelin-800-million/ ... according to a new report from the Mirror, Plant ripped up a contract promising Led Zeppelin ...</p>		

Figure 1: The interface of our proposed agreement-aware search framework, Maester. Instead of a traditional ranked list of *related* articles, we propose to present 3 *agree* articles, 3 *disagree* articles, and 5 *discuss* articles respectively for a given investigative question.

1 INTRODUCTION

Increasing amounts of rumor news have been generated and widely spread in recent years, in order to attract readership, influence opinion, and increase click-through revenue. This is a serious problem for the news industry as unreliable news increases mistrust of the media and may have wide-reaching implications such as impact on elections [5, 21]. According to a research poll, 64% of US adults say that rumor news has caused a “great deal of confusion” about the factual content of reported current events [3]. Therefore, tools for investigating rumor news have become an urgent necessity.

One useful function for such tools is to see ways a specific topic or event is represented by presenting different points of view from multiple sources. Often, these topics can be phrased as *investigative questions* such as our running example, “Did Robert Plant turn

down a contract to tour with Led Zeppelin?" For this question, some news articles reported Robert Plant turned down the contract while others disputed that it was not true; yet others merely summarized an existing article without stating its own position. In this sense, this question could be considered *controversial*. Such function is beneficial to not only users but also specialists like a journalist working on a fact-checking article or a historian cataloging beliefs and trends.

In this paper, we study how to automatically identify the stances of news articles and rank them based on their levels of agreement with a given question. Specifically, we propose Maester, a novel agreement-aware search framework. Given an investigative question, Maester will first retrieve *related* articles that address the target question. Each of these articles is then automatically assigned a stance label of either *agree*, *disagree*, or *discuss*, where *discuss* pertains to articles that merely discuss or summarize other articles reporting on the reference question without making a statement of their own with regard to the question. Splitting the results into these three categories allows the user to (a) see quickly whether a topic is controversial (e.g., some category does not have any assigned articles), (b) get an overview of the different points of view, and (c) form a more informed understanding about the sources taking a position and evidence presented in the articles.

Our methodology is based on the following two observations from real-world rumor news articles: (1) relatedness of an article can often be determined by its shared keywords/entities with the investigative question; and (2) agreement level of an article can often be inferred from a few key sentences in it. For example, as shown in Figure 1, all retrieved articles are related through the keywords "Robert Plant" and "Led Zeppelin", and we can determine their stances based on the sentences shown in the search result snippets. Accordingly, we design Maester as a two-step framework, which first filters unrelated articles and then predicts agreement status of remaining related articles. We learn a gradient boosting tree model with four types of features, including the key entity features, to classify whether an article is related to question or not. Then, we select top-3 sentences in each *related* article that are closely correlated to the investigative question. These sentences, together with the reference question, are then fed into a recurrent neural network (RNN) which outputs the level of agreement for each news article. Finally, Maester ranks these news articles and displays top-ranked ones within each agreement category to users.

We evaluate Maester using the dataset from the Fake News Challenge¹ (FNC). Extensive experiments verify our two observations empirically and demonstrate the significant improvements of Maester over the original challenge winner's solution (i.e., an ensemble model of gradient boosting trees and a convolutional neural network). In summary, our contributions are as follows.

- **Agreement-Aware Search Framework.** We propose and build a novel agreement-aware search framework, Maester, to bring a holistic view to the user towards the investigative question.
- **Agreement Detection.** We propose a novel model based on RNN with attention mechanism for classifying and ranking related articles by stance.

¹<http://www.fakenewschallenge.org/>

- **Extensive Evaluation.** We conduct a thorough experimental evaluation to demonstrate the effectiveness of Maester by comparing it with the FNC first-place method. For controversial questions, Maester achieves a significant improvement for overall agreement-aware ranking ($\sim 2x$), with a 7-fold improvement in the especially difficult case of disagreement; over both controversial and non-controversial questions, the improvement is 20%. In addition, it improves over the first-place method in terms of the FNC weighted accuracy metric by 2.88%.

2 RELATED WORK

In this section, we review literature related to agreement detection of news articles, question answering, and other lines of work relevant to our studied problem.

Stance Detection. The natural language processing community has explored stance detection for years and have formulated it in various ways. *SemEval 2016 Task 6* defines it as determining from text whether the author is in favor of, against, or neutral towards a given target [11]. In this shared task, the text is a tweet and the target is a single entity without any descriptive text. Following the same line of work, researchers have explored how to decide whether a tweet or an article favors one specific entity over others [20]. However, finding agreement with respect to an investigative question is more challenging than simply determining the stance for specific entities. This is because any subtle changes in the wording may lead to a completely different interpretation of the question.

Mohammad et al. first released a dataset for tweet stance [10], and later studied sentiment and stance for tweets [12]. Other approaches to stance detection in social media include semi-supervised topic models to classify stance [25] and latent feature extraction [27]. Furthermore, stance detection has been explored in Chinese microblogs [26] and online discussion forums [19]. All of these tasks require exactly one targeted entity, however, investigative questions may contain more than one entity. Thus, these methods cannot be directly adopted for our use case.

Agreement Detection in FNC-1. In the summer of 2017, the *Fake News Challenge* (FNC) ran its first contest on agreement detection. The task of this contest was to determine agreement given pairs of headlines and news articles. The challenge provides a partially labeled dataset, denoted in the following as *FNC-1*, which is based on the *Emergent* dataset [9], and contains rumor news. The winner of the FNC-1 [13] developed an ensemble model of a tree-based model and a CNN-based model. Similar to the solution to rumor news detection proposed in this work, the tree-based model utilizes a set of handcrafted features, however, it neglects important entity features. The CNN-based model on the other hand can extract features automatically but its performance is not as good as that of the tree-based model. We use the FNC-1 dataset for our evaluation and compare Maester with the winner's solution in Section 5 thoroughly. Note that all challenge winners [13, 24, 28] in SemEval and FNC take advantage of both handcrafted and neural network based features. Maester also follows the same paradigm.

Textual Entailment. Another related line of work is textual entailment, which studies whether a text entails, contradicts, or not related to a certain hypothesis [2, 16, 23]. However, entailment emphasizes the logical relation of text and hypothesis where the

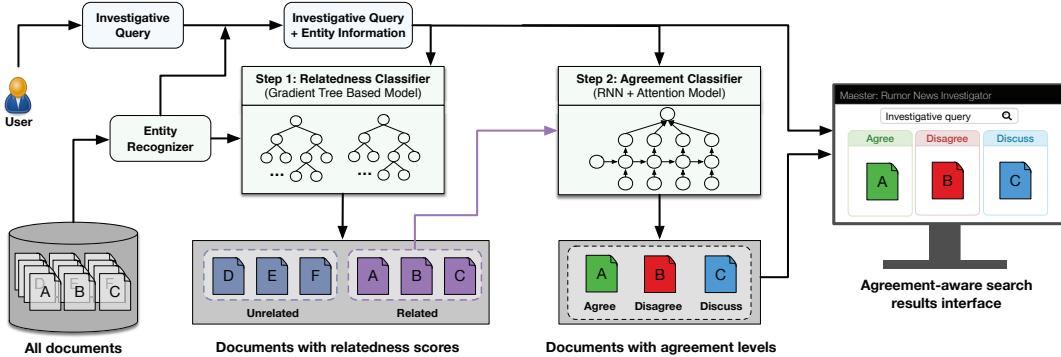


Figure 2: Overview of Maester framework.

text is commonly only one sentence and thus is much shorter than a news article.

Question Answering. Question answering (QA) is the task of finding an article, a passage, or a sentence to answer a given question [22]. Most, if not all, of these questions have a specific and clear answer. However, this work focuses on controversial questions for which traditional question answering systems do not work well. For example, given a simple fact-seeking question like “Was George Washington a U.S. president?” one should only find *agree* articles. In contrast, controversial questions lack consensus and often become a hotbed for spreading rumor news.² As a result, traditional QA systems struggle to address this modified problem.

Search Result Diversification. Search result diversification [7] has been originally proposed to deal with query ambiguity, and has been applied to improve personalized search [15] afterwards. In the same context, query reformulation [17] has been explored to retrieve more relevant articles per target, and thus diversifying the search results. In [6], the authors furthermore propose to consider the proportionality of articles instead of emphasizing diversity. However, depending on the diversity measure, articles within the same agreement group can also be diverse. Therefore, directly applying search diversification methods cannot guarantee the presence of all agreement groups. As showing multiple ranked lists for different agreement groups essentially enforces the results to be diversified, we may also apply similar techniques to optimize the overall quality of the ranked lists per agreement group.

3 PRELIMINARIES

In this section, we will first formulate the problem and then discuss our framework design and alternative models.

3.1 Problem Formulation

Given a question q , we assume that a collection of candidate articles $\mathcal{D}(q)$ is provided. There are many ways to obtain such a collection (e.g., taking the top-100 articles from a collection based on BM25 scores), which is not the focus of this paper.

DEFINITION 1 (AGREEMENT CLASSES). *Given an investigative question q and an article $d \in \mathcal{D}(q)$, we define four possible classes to describe how d relates to q :*

² We recognize the sensitivity and importance of not propagating conspiracy theories (e.g., “Did 9/11 really happen?”) and, for now, propose to deal with this challenge by limiting candidate results to trusted sources.

- (1) **Agree:** *The article agrees with q*
- (2) **Disagree:** *The article disagrees with q*
- (3) **Discuss:** *The article discusses the same question, but does not take a position w.r.t. q*
- (4) **Unrelated:** *The article addresses a question other than q .*

Previously, we have noted that the key to rumor detection is to find those questions that lead to controversial discussion of a topic, i.e., on which people have more than one opinion. More formally, we use the following definition for controversial questions.

DEFINITION 2 (CONTROVERSIAL QUESTION). *When an investigative question has at least one agreeing and one disagreeing news article in $\mathcal{D}(q)$, we refer to it as a controversial question.*

For understanding controversial questions and agreement classes, consider the following example taken from the FNC that shows text snippets referencing the running example question “Did Robert Plant turn down a contract to tour with Led Zeppelin?”. Here, the controversial question leads to different news articles that can be categorized according to statements made in those articles.

EXAMPLE 1. *The running example showing relatedness classification and agreement detection for question “Did Robert Plant turn down a contract to tour with Led Zeppelin?”*

Question	Did Robert Plant turn down a contract to tour with Led Zeppelin?
Agree	... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. . . .
Disagree	... No, Robert Plant did not rip up an \$800 million deal to get Led Zeppelin back together. . . .
Discuss	... Robert Plant reportedly tore up an \$800 million Led Zeppelin reunion deal. . . .
Unrelated	... Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today. . . .

Formal Problem Definition. Our goal is to declare whether a candidate news article is related to an investigative question and, if so, how it is positioned w.r.t. that question. More formally, we say that $\forall q \in Q$ and $d \in \mathcal{D}(q)$, there is a label $y \in \{\text{unrelated}, \text{discuss}, \text{agree}, \text{disagree}\}$ that describes the relationship between q and d . Note that it is possible that, for a given reference question, any agreement class may contain multiple news articles. Therefore, we desire the output of the agreement identification step to

Table 1: FNC-1 Dataset Statistics.

Investigative Questions			News Articles			Labeled Pairs		
All	Controversial	Total	Total	Unrelated	Discuss	Agree	Disagree	
Training	1,648	260	1,683	49,972	73.13%	17.83%	7.36%	1.68%
Testing	894	211	904	25,413	72.20%	17.57%	7.49%	2.74%

be ranked lists per class as shown in Figure 1, with k_{agree} *agree* articles, k_{disagree} *disagree* articles, and k_{discuss} *discuss* articles, for example, $(k_{\text{agree}}, k_{\text{disagree}}, k_{\text{discuss}}) = (3, 3, 5)$ as shown in the running example. To measure whether an article is related or unrelated, we determine a confidence score $\text{rel}(q, d) \in [0, 1]$ where a 0 signifies that q and d are unrelated and 1 that d is highly related to q . For related articles, their levels of agreement can be predicted by a classifier that maps an agreement score $\beta(q, d)$ to range from -1 to +1. Here -1 indicates maximum disagreement and +1 indicates maximum agreement. Our models then estimate $P(y|q, d)$ for ranking, where (1) $P(y|q, d) = \beta(q, d)$ holds for agreeing articles, (2) $P(y|q, d) = -\beta(q, d)$ holds for disagreeing articles, and (3) $P(y|q, d) = \text{rel}(q, d)$ holds for discussing articles. For each $d \in \mathcal{D}(q)$, we define its agreement \hat{y} as $\arg \max_y P(y|q, d)$. Thus, \hat{y} and the corresponding $P(\hat{y}|q, d)$ determine the membership and ranking of an article d w.r.t. q in these three lists.

Model Training & Evaluation. To train our models, we use a training set containing labels for question-article pairs as labeled above. After the models have been trained, they are evaluated on a separate set of questions and their candidate articles, as same as the training and verification methodology applied in the FNC. This process holds for both, classification and ranking, tasks.

3.2 Framework Overview

Figure 2 presents an overview of our proposed Maester framework. We structure our approach in two steps analogous to the two problems discussed above, i.e., (1) whether an article is *related* to a given question; and (2) predicting a related article’s agreement w.r.t. the question. Intuitively, the actual modeling challenges for these two problems are substantially different. We observe that content words and entity mentions in both the given question and the article may play important roles in predicting their relatedness. That is, if the article discusses the same or similar set of entities, they should be related.

OBSERVATION 1 (RELATEDNESS: KEYWORDS AND ENTITIES). *Overlapping keywords and entities between the given question q and a news article d are crucial for determining their relatedness.*

In contrast, overlapping entities are weak signals for finding the level of agreement w.r.t. a question. Specifically, either an *agree* article or a *disagree* article might contain a large number of overlapping keywords and entities. Instead, for the task of agreement detection, non-entity words such as adjective, adverbs, and negation words are more important. Furthermore, inspired by many examples such as Figure 1 and the running example in Section 3, we observe that only a few sentences, referred as *key sentences*, in an article will often reflect the stance w.r.t. a given question, especially for news articles. For example, from the sentence “No, Robert Plant did not rip up an \$800 million deal to get Led Zeppelin back together.” one

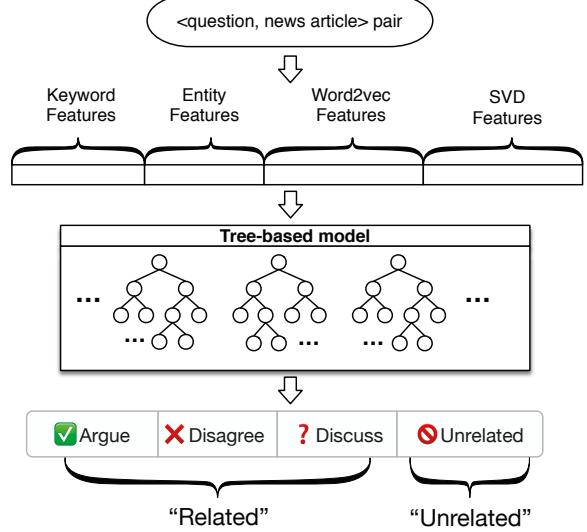


Figure 3: Tree-based Classification.

can easily derive that this article disagrees with the question “Did Robert Plant turn down a contract to tour with Led Zeppelin?”. Thus, we propose our second observation as follows.

OBSERVATION 2 (AGREEMENT: KEY SENTENCES). *An article’s agreement w.r.t. a given question q is largely decided based on a few key sentences. This is due to the “inverted pyramid” structure that journalists often follow when writing a news story [14].*

Finally, we observe that in practice, the distribution of agreement labels is often skewed. As shown in Table 1 for the FNC-1 dataset, the majority of labels are *unrelated* whereas *disagree* has the least number of annotations. Avoiding overemphasis of *unrelated* news articles further motivates the following two-step framework.

- (1) **Relatedness Classification.** First, we merge the four stances into two categories, i.e., *related* and *unrelated*, and focus on the binary classification. Based on Observation 1, for a given question and an article, we design keyword, entity, word2vec, and SVD features based on the keywords and entity mentions. Taking these features as input, as shown in Figure 3, our tree-based model leads to a test accuracy close to 98% in our experiments, which verifies this observation empirically.
- (2) **Agreement Detection.** Second, for all *related* articles, we build a 3-class classification model to estimate the agreement class. Inspired by Observation 2, for a given question and an article, we project the question and every sentence of the article into the embedding space and then choose the most similar sentences as *key sentences*. Afterwards, we inject these sentences into an efficient RNN model with attention mechanism. Note that if we instead train a tree-based model using the same

keyword/entity-based handcrafted features designed for relatedness classification, the performance drops significantly which is consistent with our observation.

4 METHODOLOGY

This section first introduces our feature design for the tree-based model which is used to compute relevance scores. Then, we present our RNN model with attention mechanism.

4.1 Relatedness Classification

In this section, we briefly introduce the features used in the relatedness classification. As shown in Figure 3, we design the following features for each question-article pair and categorize them into four different types: (1) keyword features, (2) entity features, (3) word2vec features, and (4) SVD features.

Keyword Features. We compute the non-stopword keyword overlap between the question q and the news article d , i.e., $|q \cap d| = \sum_{w \in q} \min\{freq(w, q), freq(w, d)\}$, where, $freq(w, q)$ and $freq(w, d)$ are the counts of words in the question q and the article d , respectively. Also, we add inverted document frequency to automatically scales down the importance of popular words. Furthermore, to make sure the computed scores are comparable across different questions, we normalize them to $[0, 1]$ by dividing $|q \cap d|$.

Entity Features. We apply the spaCy³ toolkit to extract named entities from questions and articles. As both question and news article may contain multiple entities, we model them using the bag-of-entities representation. Analogous to the keyword features above, we can then compute their overlaps.

word2vec Features. We utilize pre-trained word2vec 300-dimension vectors⁴ and use the average vector to build vector representations for each question and news article.

SVD Features. As an approximation, we use PCA analysis [8] to determine the topics. More specifically, we first get the TF-IDF weighted bag-of-words representations of all articles after which we apply SVD decomposition to get the principal components. Finally, we project all questions and articles onto these components to get dense feature vectors. We further compute similarity based on these dense feature vectors, which indicates whether the news articles is related to the headline or not.

Although we use similar features as the FNC winner (i.e., entity features are added and sentiment features are removed), we have achieved a substantially better classification results. More than 30% error reductions are observed in the relatedness classification in Section 5.4, which demonstrates the importance of our newly designed *entity features* based on Observation 1.

4.2 Agreement Detection

In this section, we present our recurrent neural network (RNN) with attentions model designed for agreement categorization and document ranking within certain category. Although keyword/entity-based features work well for relevance classification, they cannot

capture more subtle expressions that indicate agreement or disagreement. Recent advances on neural networks provide an automatic, high-quality way for this type of feature extraction. We design a RNN with attentions model for this purpose.

While there are many variations of long-short term memory (LSTM), we use the following one for the rumor detection problem. Suppose the input sequence is $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$, where $\mathbf{x}_k \in \mathbb{R}^l$ is the vector representation of the k -th element. At each position k , there is a set of internal vectors, including an input gate \mathbf{i}_k , a forget gate \mathbf{f}_k , an output gate \mathbf{o}_k , and a memory cell \mathbf{c}_k . All these vectors together are used to generate a hidden state $\mathbf{h}_k \in \mathbb{R}^d$ as

$$\begin{aligned}\mathbf{i}_k &= \sigma(\mathbf{W}^i \mathbf{x}_k + \mathbf{V}^i \mathbf{h}_{k-1} + \mathbf{b}^i) \\ \mathbf{f}_k &= \sigma(\mathbf{W}^f \mathbf{x}_k + \mathbf{V}^f \mathbf{h}_{k-1} + \mathbf{b}^f) \\ \mathbf{o}_k &= \sigma(\mathbf{W}^o \mathbf{x}_k + \mathbf{V}^o \mathbf{h}_{k-1} + \mathbf{b}^o) \\ \mathbf{c}_k &= \mathbf{f}_k \odot \mathbf{c}_{k-1} + \mathbf{i}_k \odot \tanh(\mathbf{W}^c \mathbf{x}_k + \mathbf{V}^c \mathbf{h}_{k-1} + \mathbf{b}^c) \\ \mathbf{h}_k &= \mathbf{o}_k \odot \tanh(\mathbf{c}_k)\end{aligned}$$

where σ is the sigmoid function, \odot is the element-wise multiplication of two vectors, and all $\mathbf{W}^* \in \mathbb{R}^{d \times l}$, $\mathbf{V}^* \in \mathbb{R}^{d \times d}$, and $\mathbf{b}^* \in \mathbb{R}^d$ are parameters to be learned.

Directly applying RNNs to model long articles is challenging. In order to capture and memorize useful information, RNNs require a bigger state size for the longer texts, and thus decrease efficiency. Fortunately, based on Observation 2, it is possible to reduce long news articles to a few key sentences with only minimal loss of output quality. To obtain these sentences, we leverage word embeddings. Considering the limited training data and the model simplicity, we define the sentence embedding as the average of its pre-trained word embeddings. Specifically, we utilize the pre-trained Glove 300-dimension vectors and skip the stopwords when computing the average vector. Since questions usually consist of one or two sentences, we apply the same approach for them. We then evaluate the cosine similarity between the given question and all sentences in a news article. The sentences with the highest similarities to the question are the key sentences which then replace the news article text. The sentences are organized in their relative similarity order. In the following, we assume a default number of key sentences k of 3. The effect of different values for k will be discussed in Section 5.7.

We follow Wang et al. [23] to build a neural attention model, as shown in Figure 4. Formally, we have two sequences $\mathbf{X}^q = \{\mathbf{x}_1^q, \mathbf{x}_2^q, \dots, \mathbf{x}_m^q\}$ and $\mathbf{X}^d = \{\mathbf{x}_1^d, \mathbf{x}_2^d, \dots, \mathbf{x}_n^d\}$, where m is the length of the question and n is the number of tokens in the selected sentences, and each \mathbf{x} is an embedding vector of the corresponding word. We build three LSTMs in total: qLSTM processes \mathbf{X}^q and generates its hidden states \mathbf{h}_k^q ; dLSTM reads \mathbf{X}^d and outputs hidden states \mathbf{h}_k^d ; and mLSTM models the matching between the question and the article and produces hidden states \mathbf{h}_k^m which we discuss in detail later.

Next, we generate the attention vectors \mathbf{a}_k ($1 \leq k \leq n$) as follows.

$$\mathbf{a}_k = \sum_{j=1}^m \alpha_{kj} \mathbf{h}_j^q \quad (1)$$

Here, α_{kj} is an attention weight that encodes the degree to which \mathbf{x}_k^d in the article is aligned with \mathbf{x}_j^q in the question.

³<http://spacy.io/>

⁴GoogleNews-vectors-negative300.bin.gz from <https://code.google.com/archive/p/word2vec/>

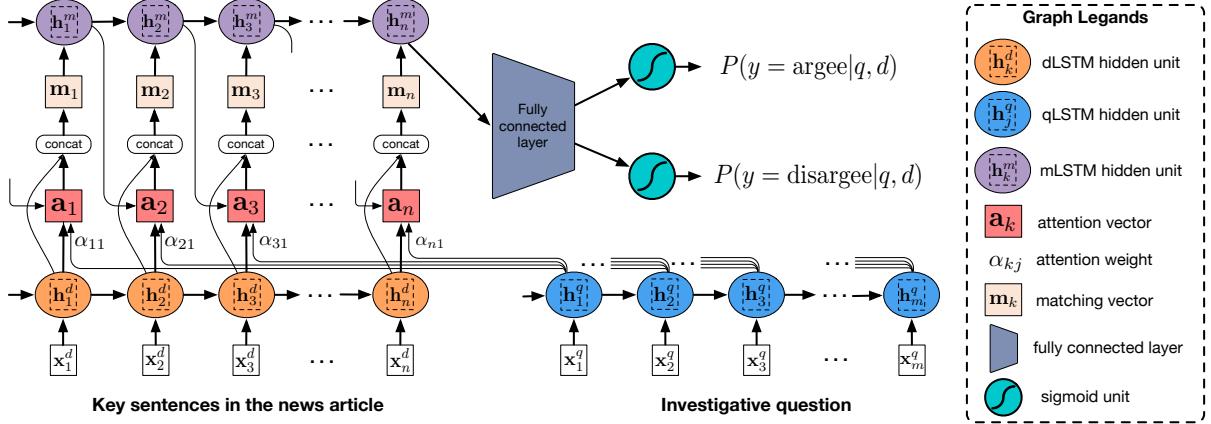


Figure 4: The architecture of our proposed RNN+attention Model.

The attention weight α_{kj} is generated as

$$\alpha_{kj} = \frac{\exp(e_{kj})}{\sum_{j'} \exp(e_{kj'})} \quad (2)$$

$$e_{kj} = \mathbf{w}^e \cdot \tanh(\mathbf{W}^q \mathbf{h}_j^q + \mathbf{W}^d \mathbf{h}_k^d + \mathbf{W}^m \mathbf{h}_{k-1}^m) \quad (3)$$

where \cdot is the dot product between two vectors and the vector $\mathbf{w}^e \in \mathbb{R}^d$ as well as all matrices $\mathbf{W}^* \in \mathbb{R}^{d \times d}$ are the parameters to be learned.

The input of mLSTM, \mathbf{m}_k , is the concatenation of \mathbf{h}_k^d , which is the hidden state for the k -th token in the article, and \mathbf{a}_k , which is its attention weighted version. Thus, mLSTM will ‘remember’ important matching results, and ‘forget’ non-essential ones.

To predict the agreement class of a news article, we use \mathbf{h}_N^m , i.e., the last hidden state of mLSTM. Instead of using a soft-max layer for 3-class classification, we choose to use two separate sigmoid modules for *agree* and *disagree*, which make the predicted scores comparable across different articles.

Furthermore, we use an agreement score $\beta(q, d) \in [-1, +1]$ with -1 indicating maximum disagreement and $+1$ indicating maximum agreement. When $\text{score}_{\text{agree}}$ is larger than $\text{score}_{\text{disagree}}$, we let $\beta(q, d)$ be a positive score of $\text{score}_{\text{agree}}$. Otherwise, we set $\beta(q, d)$ as a negative score of $-\text{score}_{\text{disagree}}$. Based on $\beta(q, d)$, we can define $P(y|q, d)$ accordingly as described in Section 3.

4.3 Online Pipeline

Once an investigative question q and its candidate collection $\mathcal{D}(q)$ arrive for processing, Maester will first apply the tree-based model to compute the relatedness score $\text{rel}(q, d)$ for each article $d \in \mathcal{D}$. Then, for the articles with $\text{rel}(q, d) \geq 0.5$, Maester will leverage the attention-based RNN to determine the agreement classes for each relevant news article. We will thus compute the agreement \hat{y} based on $P(y|q, d)$. Note that at this stage, $P(y = \text{discuss}|q, d) = \text{rel}(q, d) \geq 0.5$. Therefore, if we finally get \hat{y} as *agree* or *disagree*, its probability will be more than 0.5. The *agree* and *disagree* articles will be ranked based on the absolute values of $\beta(q, d)$, while *discuss* articles will be ranked by their $\text{rel}(q, d)$ scores.

5 EXPERIMENTS

Here we report the evaluation of Maester on the real-world dataset.

5.1 Dataset

We evaluate Maester on a recently published dataset, FNC-1⁵, from the Fake News Challenge. FNC-1 was designed as a stance detection dataset and it contains 75,385 labeled headline and article pairs. The labels are analogous to the agreement classes that we consider, namely *agree*, *disagree*, *discuss*, and *unrelated*. Each headline in the dataset is phrased as a statement. Note that our techniques hold for statements as well as investigative questions. In fact, we observe that investigative questions are most commonly rephrased statements. Detailed statistics of the dataset can be found in Table 1.

Note furthermore that the topics mentioned in the questions and articles in the training and testing sets are significantly different. Consequently, this setting is challenging and even harder than a real-world setup where partial overlap can often be assumed.

5.2 Evaluation Metrics

Since some of the questions in this dataset are not controversial, we present evaluation results in two folds: (1) all questions and (2) controversial questions. For both, we evaluate all compared methods using the following three metrics: (1) NDCG@K and Avg. NDCG for the ranking accuracy, (2) relatedness accuracy for the classifier’s performance, and (3) the official FNC metric, weighted accuracy. Considering the Maester’s interface as shown in Figure 1, we think the NDCG@K and Avg. NDCG is the most important. Details are as follows.

NDCG@K and Avg. NDCG. Because we are presenting three ranked lists of articles to the user, we utilize the normalized discounted cumulative gain, NDCG@K, for each investigative question and calculate the average over all questions for evaluation.

The gain of an article in a ranked list is defined as follows. In the ranked list of label *agree*, only *agree* articles will receive a score of 1, while other articles will get a zero score. Articles in the *disagree* and *discuss* list are treated analogously.

⁵<https://github.com/FakeNewsChallenge/fnc-1>

Given a question and a ranked list of K articles, the discounted cumulative gain is calculated as

$$\text{DCG}@K = \text{gain}_1 + \sum_{i=2}^K \frac{\text{gain}_i}{\log_2(i)}$$

The NDCG@ K is then computed as a normalization by the best possible DCG@ K . If the ideal DCG@ K is 0 for any of the lists, we will skip this ranked list for this question. Considering the numbers of articles from each class displayed in our proposed interface (i.e., Figure 1), we evaluate NDCG@3 for both *agree* and *disagree* ranked lists, and NDCG@5 for the *discuss* ranked list.

Since all questions as well as their three ranked lists are equally important for presenting the holistic view towards the investigated question to the user, to conduct an overall comparison, we define the average NDCG score as follows. For each question, we first average NDCG scores of all three ranked list. **Avg. NDCG** is computed as the average of these averages for different questions.

Relatedness Error. To evaluate the relatedness classifier, we consider only two classes: *related* vs. *unrelated*. The relatedness error refers to the percentage of misclassified question-article pairs.

Weighted Accuracy. This is the official metric for FNC-1: For a question and an article, if the model successfully predicts the *related/unrelated* label, it receives a score of 0.25. For a question and a *related* article, if the model successfully predicts *agree*, *disagree*, or *discuss*, it receives a score of 0.75. The final score is then normalized by the maximum possible score⁶.

5.3 Experimental Setting

All experiments are conducted on a single machine equipped with an Intel Xeon processor E5-2650@2.2GHz and a NVIDIA GeForce GTX 1080. In Maester, the tree-based model is implemented in XGBoost [4] and the RNN+attention model is implemented using Tensorflow [1]. The source code is available in the authors' GitHub⁷.

Maester. This is our proposed model. By default, the number of key sentences, k , is set to 3, and the number of training epochs is set to 10. For further details on the parameters, please refer to the study on parameter sensitivities in Section 5.7. As our models contain some randomness, we run all experiments five times and report the average performance.

FNC-1 Winner. As we discussed before, the FNC-1 winner's solution is an ensemble of a tree-based and a convolutional neural network (CNN) models. This combined model is able to detect the relatedness of the article effectively, primarily due to their effective tree-based model with human designed features like TF-IDF weighted keywords. However, it is limited in detecting the actual *agree* or *disagree* label of articles. Since the dataset is imbalanced, most of the related articles are labelled *discuss* and *disagree* labels are rare. Thus, the winner's solution will aggressively classify most of articles as *discuss* and the rest as *agree*, in order to achieve a high overall accuracy. However, this leads to a poor ranking performance. We report the best performance for FNC-1 Winner during the competition.

⁶For more details, please refer to <http://www.fakenewschallenge.org/>

⁷<https://github.com/shangjingbo1226/Maester>

Table 2: Error rate of relatedness classification. More than 30% of error reductions are achieved by Maester over FNC-1 Winner.

Method	All Questions	Controversial Questions
FNC-1 Winner	3.04%	3.75%
Maester	2.13%	2.46%

Alternative Models. As an alternative to our two-step framework, we also considered more straightforward models that have been applied in similar use cases before. The first of these is **bag-of-words**. It is unsuitable for our use case as language is evolving and there may be different vocabulary present in the application than in the training data. However, combining bag-of-words with some feature selection techniques leads to some interesting keywords that signal different types of agreement. For example, we observe that "reportedly" is a strong signal for *discuss*. We tried incorporating keyword lists based on the bag-of-words model in our own framework, however, improvements were negligible. Another type of models that is widely adopted when learning to match questions and articles is **matrix factorization** [18]. In our experiments, we observed that this technique has worse and unstable performance for this particular problem. Again, this is caused by the fact that not all words appearing in the application or test dataset are covered in the training data. For example, the weighted accuracy of the bag-of-words model is only 77.64%. The weighted accuracy of the matrix factorization approach is similar. Therefore, they are not included in this evaluation.

5.4 Relatedness Error

We first study Maester's performance on the relatedness classification task. As shown in Table 2. Maester has the best performance and achieves more than 29.93% and 34.40% error reductions on all questions and controversial questions, respectively. This demonstrates the importance of the added entity features compared to previously utilized sentiment features which tend to be noisy. An error rate less than 3% demonstrates that Maester's tree-based model built upon handcrafted features is precise enough to predict whether a document is related or not.

To compare the significance of different features, we calculate the relative feature importance for each feature type using the built-in function in XGBoost [4], as shown in Table 3. Here, we can see that the combined importance of keyword features and entity features is significant, i.e., 52.18%. Moreover, the newly added entity features are more important than the word2vec and SVD features. Therefore, Observation 1 has been verified with this experiment.

5.5 Ranking Evaluation

We evaluate the results as three ranked lists. This ranking evaluation is crucial because our ultimate goal is to present a holistic view towards the user's question.

As shown in Table 4, Maester achieves the best overall agreement-aware ranking performance. Maester's Avg. NDCG score is much higher than FNC-1 Winner's Avg. NDCG score, for both controversial and non-controversial questions. Specifically, for controversial questions, Maester's almost doubles FNC-1 Winner's performance, while for both controversial and non-controversial questions, the

Table 3: Feature importance.

Feature	Importance
Keyword	29.68%
Entity	22.50%
word2vec	13.75%
SVD	34.07%

Table 5: Weighted accuracy of agreement detection. Note that FNC-1 winner wins the challenge by an advantage of 0.05%. Maester’s improvements should be considered as remarkable.

Method	All Questions	Controversial Questions
FNC-1 Winner	82.02%	66.66%
Maester	82.98%	69.54%

improvement is 20%. We also notice that *disagreement* class is the most challenging one among all the three classes, and Maester achieves a 7-fold improvement for this class.

The improvements on the NDCG score in the *discuss* class are also noticeable. The NDCG score in the *agree* class is slightly lower than the reference score but is still comparable. These significant ranking improvements demonstrate that Maester is a better fit than FNC-Winner as a helpful rumor news investigation tool.

Finally, from this ranking evaluation, we obtain a better understanding about the FNC-1 Winner. It achieves the high weighted accuracy through aggressively predicting articles as *agree* and *discuss* where very few articles are categorized as *disagree*. However, such biased prediction gets punished when evaluating ranking performance.

5.6 FNC metric: Weighted Accuracy

Since FNC-1 Winner is specifically optimized for the official metric (i.e., weighted accuracy) in the challenge, we also used the weighed accuracy for evaluation. From Table 5, we can find that Maester outperforms FNC-1 winner where the absolute improvement of accuracy is 0.96% and 2.88% on all questions and controversial questions, respectively. Considering that FNC-1 winner has won the FNC by a margin of 0.05%, these improvements can be considered as remarkable.

In fact, recall that Maester relies only on the top-3 key sentences from the article, whereas FNC-1 Winner considers all sentences in the article. These results reflect that using only three key sentences can still capture enough information to detect agreement.

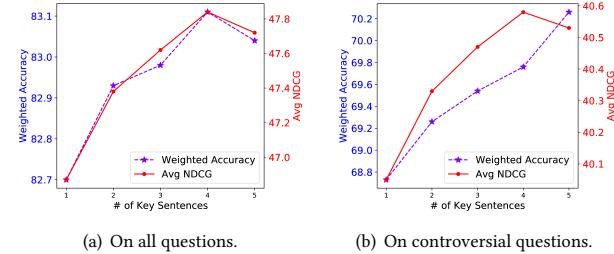
5.7 Parameter Sensitivities

Here, we study the parameter sensitivities for the two major parameters in Maester: (1) the number of key sentences, k and (2) the number of epochs needed for model convergence.

As shown in Figure 5 only knowing the top sentence of an article already provides good quality results. When more key sentences are available, the weighted accuracy on controversial questions grows constantly, while the ranking performance drops a little when $k = 5$ is reached. This implies that more sentences disclose more information, however, a few key sentences are enough for good ranking quality, which supports Observation 2.

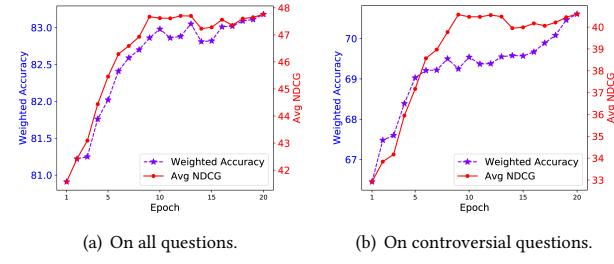
Table 4: Ranking performance of the agreement-aware search framework.

Method	All Questions			Controversial Questions				
	Agree NDCG@3	Disagree NDCG@3	Discuss NDCG@5	Avg. NDCG	Agree NDCG@3	Disagree NDCG@3	Discuss NDCG@5	Avg. NDCG
FNC-1 Winner	51.71%	2.31%	64.04%	39.38%	43.75%	2.58%	31.90%	26.08%
Maester	48.11%	20.38%	68.20%	47.62%	40.88%	19.13%	61.39%	40.47%



(a) On all questions.

(b) On controversial questions.

Figure 5: How many key sentences are enough?

(a) On all questions.

(b) On controversial questions.

Figure 6: Convergence study on test data.

Second, we studied the convergence of the RNN+attention model in Maester in Figure 6. The results show that the result quality, measured with either weighted accuracy or Avg NDCG, stabilizes after 10 epochs. This is a promising time span for early stops and savings on training time.

5.8 Efficiency Evaluation

The whole Maester pipeline, including both tree-based and RNN models, can be trained within 1 hour. However, in a real-world application, online serving time is more important. Maester can process a pair of question and article within about 5.86 ms. Specifically, in our setup, Maester spends about 0.16 seconds on average to present the final results (as shown in Figure 1) to the user.

5.9 Case Study

For a controversial question, we randomly pick two articles from the *agree* and *disagree* classes and show the top-3 key sentences selected by Maester in Table 6. From these results, we observe that the chosen sentences, especially the highlighted parts, are essential for agreement classification. Moreover, for this question, Maester achieves 100% NDCG@3 in both *agree* and *disagree* ranked lists, while the FNC-1 winner’s scores are 29.82% and 0%, respectively. These findings further consolidate our Observation 2.

6 CONCLUSION & FUTURE WORK

In this paper, we focus on investing rumor news using an agreement-aware article search. We develop an agreement-aware search framework that is designed to provide users with a holistic view of an

Table 6: Top-3 key sentences determined by Maester for agreement detection.

Question	Is it true that a woman pays \$20,000 for third breast to make herself LESS attractive to men?
An agree article	<ol style="list-style-type: none"> 1. No, you do not need to adjust your sets, you are actually looking at a woman with three breasts. 2. Jasmine added: I got it because I wanted to make myself unattractive to men. 3. She denies that she had the extra breast put on to get fame and fortune.
A disagree article	<ol style="list-style-type: none"> 1. Did a woman claiming to have a third breast play a hoax on us? 2. A top plastic surgeon, Mr Nilesh Sojitra, also cast doubt over the surgery after claiming no reasonable doctor would perform the operation. 3. Snopes.com came up with a number of intriguing arguments that could indicate Jasmine Tridevil did not actually pay \$20,000 for an extra breast.

investigative question, for which the ground truth is not certain. Based on two intuitive but important observations, we designed a two-step model consisting of a tree-based model based on hand-crafted features and a RNN+attention model focusing on only a few key sentences. Our experimental results and case studies not only demonstrate the effectiveness of our model, but also verify both observations empirically.

There are many related problems and follow-up work that should be explored in the future. In the context of rumor detection, we propose using statements, here in the form of controversial questions, to further the understanding of a topic. However, it remains unclear how to derive such statements. Another line of interesting follow-up work is to allow not only a limited set of labels but to enable additional entity-driven options. For example, given the question “Who is the best basketball player in history?” many people will say “Michael Jordan” but there are others who will mention names such as “Kobe Bryant” and “Lebron James”.

ACKNOWLEDGEMENTS

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HD-TRA11810026, Google PhD Fellowship, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

- [1] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEAWAT, S., IRVING, G., ISARD, M., ET AL. Tensorflow: A system for large-scale machine learning. In *OSDI* (2016), vol. 16, pp. 265–283.
- [2] ANDROUTSOPoulos, I., AND MALAKOSIOTIS, P. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38 (2010), 135–187.
- [3] BARTHEL, M., MITCHELL, A., AND HOLCOMB, J. Many americans believe fake news is sowing confusion. *Pew Research Center* 15 (2016).
- [4] CHEN, T., AND GUESTRIN, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), ACM, pp. 785–794.
- [5] CONNOLLY, K., CHRISAFIS, A., MCPHERSON, P., KIRCHGAESSNER, S., HAAS, B., PHILLIPS, D., HUNT, E., AND SAFI, M. Fake news: an insidious trend that's fast becoming a global problem. *The Guardian* 2 (2016).
- [6] DANG, V., AND CROFT, W. B. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (2012), ACM, pp. 65–74.
- [7] DROSOU, M., AND PITOURA, E. Search result diversification. *ACM SIGMOD Record* 39, 1 (2010), 41–47.
- [8] DUNTEMAN, G. H. *Principal components analysis*. No. 69. Sage, 1989.
- [9] FERREIRA, W., AND VLACHOS, A. Emergent: a novel data-set for stance classification. In *HLT-NAACL* (2016).
- [10] MOHAMMAD, S., KIRITCHENKO, S., SOBHANI, P., ZHU, X.-D., AND CHERRY, C. A dataset for detecting stance in tweets. In *LREC* (2016).
- [11] MOHAMMAD, S., KIRITCHENKO, S., SOBHANI, P., ZHU, X.-D., AND CHERRY, C. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval@NAACL-HLT* (2016).
- [12] MOHAMMAD, S. M., SOBHANI, P., AND KIRITCHENKO, S. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17, 3 (2017).
- [13] PAN, Y., SIBLEY, D., AND BAIRD, S. Fake news challenge - team solat in the swen. <https://github.com/Cisco-Talos/fnc-1>, June 2017. Accessed: 2017-12-27.
- [14] PO' TTKER, H. News and its communicative quality: The inverted pyramid. In *Journalism Studies* 4, 4 (2003), 501–511.
- [15] RADLINSKI, F., AND DUMAIS, S. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), ACM, pp. 691–692.
- [16] ROCKTÄSCHEL, T., GREFENSTETTE, E., HERMANN, K. M., KOČISKÝ, T., AND BLUNSOM, P. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* (2015).
- [17] SANTOS, R. L., MACDONALD, C., AND OUNIS, I. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 881–890.
- [18] SHANG, J., CHEN, T., LI, H., LU, Z., AND YU, Y. A parallel and efficient algorithm for learning to match. In *Data Mining (ICDM), 2014 IEEE International Conference on* (2014), IEEE, pp. 971–976.
- [19] SKEPPSTEDT, M., KERREN, A., AND STEDE, M. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)* (2017), pp. 1–8.
- [20] SOMASUNDARAN, S., AND WIEBE, J. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (2009), Association for Computational Linguistics, pp. 226–234.
- [21] TAVERNISE, S. As fake news spreads lies, more readers shrug at the truth. *The New York Times* (2016).
- [22] VOORHEES, E. M., ET AL. The trec-8 question answering track report. In *Trec* (1999), vol. 99, pp. 77–82.
- [23] WANG, S., AND JIANG, J. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849* (2015).
- [24] WEI, W., ZHANG, X., LIU, X., CHEN, W., AND WANG, T. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *SemEval@ NAACL-HLT* (2016), pp. 384–388.
- [25] XU, K., BI, S., AND QI, G. Semi-supervised stance-topic model for stance classification on social media. In *Joint International Semantic Technology Conference* (2017), Springer, pp. 199–214.
- [26] XU, R., ZHOU, Y., WU, D., GUI, L., DU, J., AND XUE, Y. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *NLPCC/ICCPOL* (2016).
- [27] XU, X., HU, F., DU, P., WANG, J., AND LI, L. Efficient stance detection with latent feature. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data* (2017), Springer, pp. 21–30.
- [28] ZARRELLA, G., AND MARSH, A. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784* (2016).