TrClaim-19: The First Collection for Turkish Check-Worthy Claim Detection with Annotator Rationales

Yavuz Selim Kartal and Mucahid Kutlu
Department of Computer Engineering
TOBB University of Economics and Technology
Ankara, Turkey
{ykartal,m.kutlu}@etu.edu.tr

Abstract

Massive misinformation spread over Internet has many negative impacts on our lives. While spreading a claim is easy, investigating its veracity is hard and time consuming, Therefore, we urgently need systems to help human fact-checkers. However, available data resources to develop effective systems are limited and the vast majority of them is for English. In this work, we introduce TrClaim-19, which is the very first labeled dataset for Turkish check-worthy claims. TrClaim-19 consists of labeled 2287 Turkish tweets with annotator rationales, enabling us to better understand the characteristics of check-worthy claims. The rationales we collected suggest that claims' topics and their possible negative impacts are the main factors affecting their check-worthiness.

1 Introduction

In 2013, World Economic Forum (WEF) listed massive digital misinformation as one of the top global risks likely to occur in 10 years[1]. Unfortunately, we witnessed many unpleasant incidents due to misinformation spread over Internet such as massive stock price changes[2], gunfights[3], and others. Since the start of COVID-19 pandemic, we have also observed many incidents showing the value of true information and how misinformation about health issues can be deadly (e.g., misusing disinfectants to prevent coronavirus after Donald Trump suggested injecting disinfectants as treatment[4]).

In order to prevent the negative outcomes of misinformation, many fact-checking websites emerged all over the world in the last decade (Cherubini and Graves, 2016). The fact-checking websites manually investigate veracity of claims and share their findings with their readers. While they play an important role in the combat against misinformation, their precious journalistic effort is not enough to reduce spread of misinformation and its negative outcomes. While making a claim is so easy, investigating its veracity is highly time consuming, taking around one day (Hassan et al., 2017). Furthermore, Vosoughi et al. (2018) report that misinformation spread eight times faster than true information. Hence, we need effective solutions to help human fact-checkers and to reduce the negative impact of misinformation.

As Nakov et al. (2018) outline, the first task of a fact-checking system is to detect whether a statement contains a check-worthy claim or not. Considering the massive amount of messages shared on social media platforms, check-worthy claim detection models help human fact-checkers to filter out unimportant claims and use their valuable time to detect veracity of the most important claims. A number of researchers worked on this problem (e.g., (Lespagnol et al., 2019; Hassan et al., 2017; Jaradat et al., 2018)) and shared tasks for check-worthy claim detection have been organized (Nakov et al., 2018; Atanasova et al., 2019; Barrón-Cedeno et al., 2020).

While researchers showed great interest in fact-checking, the available resources are still limited and the vast majority of the studies focused on English. Regarding the task of detecting check-worthy claims, the only available labeled datasets are for English and Arabic (Nakov et al., 2018; Atanasova et al., 2019). However, as WEF notes in its aforementioned report, misinformation is a global problem affecting all countries. Misinformation can

[1] http://reports.weforum.org/global-risks-2013

[2] www.reuters.com/article/net-us-usa-whitehouse-ap/hackers-send-fake-market-moving-ap-tweet-on-white-house-explosions-idUSBRE93M12Y20130423

[3] www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[4] https://www.reuters.com/article/us-health-coronavirus-disinfectants-idUSKBN23C2P2

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 386–395

https://doi.org/10.18653/v1/P17

also spread internationally. For instance, during 2019 European elections, same or similar stories have been shared in different languages across European countries (Fletcher et al., 2018). Hence, in order to have an effective combat against spread of misinformation, we need research studies for a wide range of languages.

In this work, we focus on Turkish and introduce TrClaim-19, which is the very *first* labeled Turkish tweets with the rationales of annotators for check-worthy claim detection task. Turkish is a particularly important language for fact-checking studies because Fletcher et al. (2018) report that 49% of Internet users in Turkey coincide with at least one fake news in a week, which is higher than all other countries investigated in their study. Furthermore, being a member of Altaic language family, Turkish language has different linguistic features than other languages studied for fact-checking, such as being an agglutinative language and having flexible word order structure in sentences. In addition to developing a useful resource for the research community, we also seek answers for the following research questions.

- RQ-1: What is the agreement level between non-expert fact-checkers on check-worthiness of claims?
- RQ-2: Do non-experts have different opinion about check-worthiness of claims than experts?
- RQ-3: What are the main rationales to label claims as check-worthy?

In particular, we have first crawled Turkish tweets for 344 days in 2019, tracking important events happened in Turkey such as local elections, earthquake in Istanbul, and military operation in Syria. Eventually, we gathered around 225 millions Turkish tweets. Subsequently, we crawled 765 claims fact-checked by two Turkish fact-checking websites. Next, for each claim, we retrieved three tweets from our tweet crawl using Lucene search engine library[5]. Each retrieved tweet has been labeled by three separate annotators. For each tweet, we asked annotators whether it is relevant to the respective claim, and whether it contains a check-worthy claim. Inspired by McDonnell et al. (2016)'s study, we also asked their rationale for the tweets labeled as check-worthy.

[5] https://lucene.apache.org/core/

Table 1: General Statistics about TrClaim-19.

The number of tweets crawled	225M
The number of tweets annotated	2287
The number of check-worthy claims	875
The number of rationale categories	26

Table 1 summarizes general features of TrClaim-19. In total, we collected labels for 2287 tweets, and 875 of them are labeled as check-worthy when labels are aggregated by majority voting. We have observed that agreement on check-worthiness of tweets among non-experts are low (Fleiss' kappa = 0.23). In 36% of cases, non-experts disagreed with experts on check-worthiness of claims. Assessors provided rationales in 26 different categories. Rationales we collect suggest that topics and possible negative impacts of claims are the main factors in making a claim check-worthy.

The contributions of our work are as follows.

- We develope and share TrClaim-19, which is the very first labeled data resource for Turkish check-worthy claim detection [6].
- TrClaim-19 is also the first data resource with annotator rationales for check-worthy claim detection task, enabling better understanding of the research problem to develop effective solutions.
- We investigate the subjectivity of check-worthiness of claims. In particular, we explore how much non-expert and expert fact-checkers agree on check-worthiness of claims.
- We provide performance results of four models on TrClaim-19 to provide reference baselines for future studies.

2 Related Work

A number of researchers created annotated datasets for check-worthy claim detection. To our knowledge, Hassan et al. (2015) created the first check worthiness dataset using transcripts of the U.S. presidential debates. They invited journalists, professors and university students to label the data. Each sentence in debates is labeled by at least two annotators. They used a three-scale label which are "non factual", "unimportant factual" and "check-worthy factual". In total, they labeled 1571 sentences.

[6] https://github.com/YSKartal/TrClaim19

Following Hassan et al. (2015), US political debates have been used in many other studies. Patwari et al. (2017) constructed a check-worthy detection dataset using US primary and presidential debates. They identified 9 reputable news portals and fact-checking websites (e.g., Polifact, CNN and factcheck.org), and considered statements which are fact-checked by at least one of these websites, as check-worthy. Gencheva et al. (2017) also applied the same method to build their dataset using again US debates. They report that the agreement between fact-checking websites is low: Only one sentence is fact-checked by all nine websites they listed while 880 statements are fact-checked by only one website. In a follow-up study, Jaradat et al. (2018) introduce ClaimRank for detecting check-worthy claims in English and Arabic. However, the Arabic data they use is just the translation of the US election debates.

CLEF Check That! Lab have been organizing shared tasks for detecting check-worthy claims since 2018. In 2018, similar to other datasets, they used US debates and political speeches, and considered statements fact-checked by FactCheck.org as check-worthy (Elsayed et al., 2018). For Arabic check-worthy detection task, they just used translations of the English dataset. In 2019, the organizers of the lab used the extension of the previous year's dataset (Atanasova et al., 2019). In 2020, CLEF Check That! Lab[7] organizers constructed datasets using tweets and again political debates (Barrón-Cedeno et al., 2020). The English tweet dataset consists of 962 tweets about COVID-19 pandemic. The Arabic dataset consists of 7500 tweets about 15 trending topics at the time of crawling among Arab social media users such as COVID-19 and Trump Peace Plan[8].

Overall, the vast majority of the existing datasets are English and use U.S. political speeches and debates. TrClaim-19 distinguishes from the existing data resources as follows. 1) Tr-Claim19 is the first collection for Turkish check-worthy claims. 2) TrClaim-19 is the first collection providing with the rationales of annotators. 3) We collect labels for each tweet from three annotators, enabling us to analyze subjectivity of the task and different rationales for the same check-worthy claims.

[7] https://sites.google.com/view/clef2020-checkthat/

[8] https://en.wikipedia.org/wiki/Trump_peace_plan

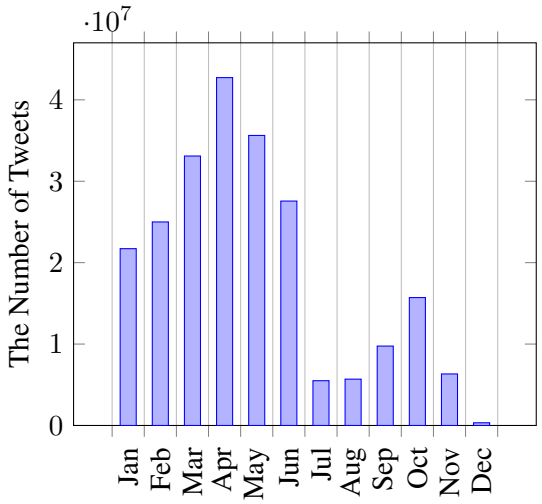

Figure 1: The number of tweets crawled in each month of 2019.

3 Tweet Collection

Social media platforms play an important role in spreading misinformation. Therefore, we focus on claims spread on Twitter. In order to gather tweets with check-worthy claims, we crawled Turkish tweets between January 1, 2019 and December 31, 2019[9] using Twitter API. We used keyword tracking approach enabling us to gather tweets about important topics. In 2019, there were many major events happened in Turkey, such as local elections[10], an earthquake in Istanbul [11], a military operation in Syria[12], and others. Therefore, we used keywords related to these major events such as "seçim" (election), "terrör" (terror), "ekonomi" (economics), "mülteci" (refugee), "Suriye" (Syria), politician names (e.g., "Erdoğan"), and others. We used a dynamic keyword list, instead of a static one, such that we updated the list whenever a new important event happened. For instance, after the earthquake in Istanbul in September 26, 2019, we added the word "deprem" (earthquake) in our keyword list.

At the end of our long crawling process, we have collected around 225 millions Turkish tweets. The distribution of tweets crawled for each month is shown in **Figure 1**. The number of tweets collected

[9] crawling had stopped only for three weeks in July 2019

[10] www.bbc.com/news/world-europe-47764393

[11] https://en.wikipedia.org/wiki/2019_Istanbul_earthquake

[12] www.aa.com.tr/en/operation-peace-spring/operation-peace-spring-starts-in-n-syria-erdogan/1607147

between January and June is much higher than other months because local elections were hold on 31^{st} March and 24^{th} of June (repeated elections for Istanbul) and Twitter was one of the main platforms for political discussions.

4 Tweet Selection

As we have collected tweets more than we can annotate, we have to select a sample among them. A random sampling could cause many not-check-worthy tweets. Inspired by test collection construction methodology in the field of information retrieval (Voorhees, 2001), we first collected claims fact-checked by fact-checking websites (Section 4.1) and then used each claim as a search query to retrieve tweets which are similar to claims we collected (Section 4.2). Using tweets similar to claims also allow us to seek answer for our research question RQ-2. Now we explain our tweet selection process in detail.

4.1 Claim Crawling

In order to ensure the quality of our collection, we focused on fact-checking websites certified by International Fact Checking Network (IFCN)[13] which investigates fact-checking websites on various issues such as nonpartisanship and transparency of fact-checking methodology. Thus, we gathered claims fact-checked by Teyit[14] and Doğruluk Payı[15] (DP) which are the only Turkish fact-checking websites certified by IFCN. Teyit usually focuses on claims spread on social media platforms while DP investigates veracity of claims made by politicians. We crawled all claims fact-checked by Teyit and DP between February 2015 and January 2020. While tweets we crawled have been posted in 2019, we did not filter claims based on their date. This is because people might make the same claim multiple times. For instance, many people believe that the moon landing in 1969 was fake and this claim circulates for decades.

We eliminated claims about pictures or videos because we focus on claims that can be analyzed linguistically. In total, we collected 573 claims from DP and 192 from Teyit, yielding 765 claims in total.

[13] www.poynter.org/ifcn/
[14] www.teyit.org
[15] www.dogrulukpayi.com

Algorithm 1 Tweet Selection

Input: Tweets T, Claims C
Output: Selected Tweets ST

1: Build an index for T
2: $ST \leftarrow []$
3: **for each** claim **in** C **do**
4: $\quad ranked \leftarrow BM25(index, claim)$
5: $\quad ST \leftarrow ST \cap ranked[0]$
6: $\quad i \leftarrow 1$
7: $\quad count \leftarrow 0$
8: $\quad$ **while** $i < ranked.size()$ **do**
9: $\quad\quad S \leftarrow LDF(ranked[i], ranked[i-1])$
10: $\quad\quad$ **if** $score < threshold$ **then**
11: $\quad\quad\quad ST \leftarrow ST \cap ranked[i]$
12: $\quad\quad\quad count \leftarrow count + 1$
13: $\quad\quad\quad$ **if** $count == 3$ **then**
14: $\quad\quad\quad\quad$ **break**
15: $\quad\quad i \leftarrow i + 1$
16: **return** ST

4.2 Selecting Tweets to be Labeled

In selection of tweets to be labeled, we have three main goals: 1) building a balanced test collection in term of label distribution, 2) having a diverse set of tweets in terms of topic and linguistics features, and 3) ensuring that some of tweets contain claims fact-checked by experts for our research question RQ-2.

Our tweet selection algorithm is presented at **Algorithm 1**. We first index all tweets without stemming using Lucene search engine library [Line 1]. This process also eliminates exactly same tweets, which exist a lot in tweet collections due to retweets. Subsequently, for each claim, we use the claim itself as a search query and rank the tweets using BM25 ranking function [Line 4]. However, instead of getting the top ranked tweets, we apply a two-step process to find tweets related to our crawled claims but different each other. This is because we observed that many tweets are not exactly same but are quite similar. For instance, two tweets might have the same message but one of them has an additional URL, hashtag, and/or emoji. In particular, we first select the top ranked tweet to be labeled [Line 5]. Then, starting from the top ranked tweets, we calculate textual similarity of tweets ranked consecutively [Line 9]. If similarity of tweets are lower than a pre-defined threshold value, we add the respective tweet (i.e., tweet with a lower rank) to our selected tweet list [Lines 13-

14]. In order to calculate textual similarity between two tweets, we first remove mentions, URLs and non-alphanumerical characters, and then calculate the approximate similarity based on Levenshtein Distance[16]. Empirically, we set the similarity score threshold to 0.80. We select three tweets for each claim using this method [Lines 12-14].

Using the algorithm explained above, we selected 1714 tweets for claims crawled from DP, and 573 tweets for claims crawled from Teyit, yielding 2287 tweets in total[17].

5 Annotation Process

Two authors of this paper and 7 volunteers annotated the tweets we selected. All annotators are native speakers of Turkish language and have degrees at various disciplines including law, computer science, dentistry, history, and business administration. Their ages are between 22 and 35. We believe that having annotators with diverse backgrounds allows us to capture different thoughts and perspectives about check-worthy claims.

Before starting the annotation process, we explained the dataset and the annotation task to each volunteer. In the annotation interface, annotators are able to see tweets to be labeled and also respective claims used to retrieve them. Annotators are asked the following three questions in the annotation task.

- Is the tweet relevant to the claim?
- Do you think that the tweet contains a check-worthy claim?
- If you think the tweet has a check-worthy claim, why do you think it is check-worthy?

We did not put any restrictions to define their rationale not to induce any bias. Annotators used a free-style text form to type their rationale behind their check-worthiness judgment. We only requested annotators to do their best to be consistent in their rationale definition and use the same text for the same rationale to better analyze their rationales in the post-process of the data.

Each tweet has been annotated by three annotators separately. We divided the participants into 3 groups. Then each annotator group labeled a separate set of tweets. The sizes of tweet sets were determined based on availability of each annotator group because they did this annotation task voluntarily. In particular, the number of tweets assigned to each group are, 384, 427, and 1476. In order to prevent bias in the annotations, the annotators have not seen labels of others until the labelling process ends. In total, we collected 2287x3=6861 relevance and check-worthiness judgments.

6 Analysis of TrClaim-19

Table 2 presents statistics about TrClaim-19. Aggregating each tweet's relevance and check-worthiness judgments based on majority voting yields 974 (=78+267+387+242) relevant (i.e., 42% of our collection) and 875 (=387+242+211+35) check-worthy claims (i.e., 38% of our collection) in total.

RQ-1 What is the agreement level between non-experts on check-worthiness of claims? The annotations we collected show that non-expert fact-checkers (i.e., annotators in our study) highly disagree on check-worthiness of claims. In Table 2, we observe that annotators fully agree on check-worthiness of 1033 (=78+242+678+35) tweets (i.e., 45% of all collection) but not for the remaining 1254 tweets (i.e., 55% of all collection). Fleiss' kappa score for check worthiness judgments is 0.23, which accounts for 'fair agreement' according to Fleiss et al. (1971). As a reference point, Fleiss' Kappa score for relevance judgments is 0.61 which accounts for 'substantial agreement'. This suggests that judging check-worthiness of claims is a more subjective task than relevance judging.

As we mention before, we had grouped annotators into three groups. In order to see whether there is any difference in annotation agreement levels among groups, we also calculated Fleiss' kappa score for each group separately. Fleiss' Kappa scores for each group's check-worthiness judgments are between 0.17 and 0.32, suggesting that there is no notable difference among groups.

Overall, regarding our research question, our results suggest that non-expert people do disagree on check-worthiness of claims. Therefore, instead of using binary judgments as in existing datasets, we believe that graded judgments are more suitable for this task. Note that TrClaim-19's judgments can be easily converted to graded judgments using check worthy ratios for each tweet.

[16] https://github.com/seatgeek/fuzzywuzzy

[17] for some claims, the search engine returned less than three tweets

Table 2: Label Distribution. Relevance judgments are aggregated based on majority voting. Check Worthy Ratio shows the ratio of "check worthy" annotations among three annotations for each tweet.

Relevance	Check Worthy Ratio	The Number of Tweets		
		Teyit	DP	Total
Relevant	0/3	32	46	78
	1/3	85	182	267
	2/3	141	246	387
	3/3	142	100	242
Non-Relevant	0/3	84	594	678
	1/3	52	337	389
	2/3	29	182	211
	3/3	8	27	35

Table 3: Sample Tweets Relevant to the Respective Claims Fact-Checked by Experts with Varying Check Worthy Judgment Ratios

Original	Translation	CW Ratio
Yüksek teknolojili ürünlerin imalattaki payı yüzde 3'e geriledi	The share of high-tech products in production fell to 3 percent	0/3
Bugüne kadar Suriyeli sığınmacılar için harcanan para 40 milyar Dolar	The total amount of money spent for Syrian refugees are 40 billions dollars.	1/3
Trafik kazalarında son on yılda 52 bin 95 kişi yaşamını yitirdi	52,095 people have died due to traffic accidents in the last ten years	2/3
Avrupa'da genç işsizliği yükselen iki ülkeden biri Türkiye	Turkey is one of the two countries in Europe in which youth unemployment increases.	3/3

RQ-2 Do non-experts have different opinion about check-worthiness of claims than experts? Relevant tweets in TrClaim-19 might be useful to seek an answer for this research question. Claims we crawled from Teyit and DP have been fact-checked by experts. Thus, we can safely assume that expert fact-checkers considered these claims as check-worthy. In Table 2, we see that check-worthy ratio of 78 relevant tweets is 0/3 (i.e., all annotators disagreed with experts). Furthermore, in 267 tweets, two-third of annotators considered the tweets as not check-worthy in contrast to experts. Overall, in 36% ($= \frac{78+267}{78+267+387+242}$) cases, annotators disagreed with expert fact-checkers.

In order to shed light this disagreement between experts and non-experts, **Table 3** shows sample relevant tweets (original Turkish tweets and their translations) for varying check-worthy ratios. We selected tweets that all annotators judged as relevant to the respective claim. In the table, we observe that none of the annotators considered the claim about the share of high-tech products in production. This might be because their life is not directly affected by this claim about economics. On the other hand, all annotators considered the claim about youth unemployment as check-worthy. This might be because the annotators are young and, therefore, interested in youth unemployment.

During our annotation process, annotators knew that the respective claims are fact-checked by experts. This might potentially affect their judgments. We leave the analysis of this potential bias as future work. However, regarding RQ-2, it is noteworthy that annotators disagreed with experts in many cases despite they knew that experts fact-checked those claims.

RQ-3: What are the main reasons to label a claim as check-worthy? Annotators judged tweets as check-worthy 2683 times (i.e., without aggregation) and provided rationales for these cases. As we did not put any restriction on how they should define rationales, they provided 71 different texts as their rationales in total. We observed that annotators used different texts for same or similar rationales (e.g., "refugee rights", "refugees", and "refugee problems"). Therefore, we manually in-

Table 4: The Most Common Rationale Groups in TrClaim-19 and a Sample Tweet for Each Group. Rationales and tweets are manually translated to English. The numbers in parantheses represent how many times rationales in the respective group appear in Tr-Claim-19.

Rationale Group	**Actual Rationales Provided by Annotators**	**A Sample Tweet**
Economics (508)	"economy", "assessment of economical status", "the impact of governmental decisions on economics"	In Turkey, 70% of the population is in debt, the poorest people get only 6.1% of GDP
Politics (376)	"security of voting", "impact on society after elections", "election time and its impacts", "politics and its impact on society", "political", "It can negatively impact governmental institutions and politicians", "political propaganda", "It might affect people's political stance", "about a political party", "municipality services"	"Republican People's Party's investigation proposal about attacking Kılıçdaroğlu has been rejected"
Society (287)	"Social problems", "gender inequality in society", "It is about an important topic for the society", "Assessment of people's psychosocial status", "About the society", "national"	"We increased women's labor force participation to 34.1%"
General (187)	"The assessment of an existing issue"	"With its 82 million population, Turkey is 18^{th} most crowded country. In addition. It has the 19^{th} largest economy in the world."
International (58)	"International Politics", "Foreign agenda", "Universal", "The impact of events happened in Turkey on other issues", "International Comparison"	United States Secretary of State Mike Pompeo: "Trump is ready to use military force against Turkey if needed."
Interesting (41)	"The claim is interesting", "strange"	"Europe Central Bank has printed Euro souvenir banknotes featuring Ataturk for collectors."
Security (35)	"Security", "military", "terror"	Yesterday it was a dream that Turkey would produce its own national tank, rifle, helicopter, submarine and UAV. Now it became a reality with Erdogan. The ratio of national defense systems (used by the army) increased from 15% to 75%.
Unclear & Suspicious Claims (33)	"Suspicious", "Checking a partial information about something"	It is said that President Erdogan spent 1,006,621 Turkish Liras of the discretionary fund in the first five months of 2019
Education (33)	"Education"	The number of unemployed people with a university degree exceeded one million
Health (28)	"health", "accident", "work accident"	Carefulness increased, work accidents decreased

vestigated all rationales and grouped them under 26 categories[18]. We observe that some of the rationales are just a few words (e.g., "human rights", "health"), suggesting that the claim is check-worthy because it is about an important topic. On the other hand, some annotators provided clear rationales

[18]We share both original rationales and our rationale groups

such as "The claim is interesting", "People might make an important decision about their life because of this claim", and "It can negatively affect the perception towards people or institutions".

Table 4 shows the most frequently stated 10 rationale groups, the actual rationale texts provided by annotators and a sample tweet for each group. The rest of the rationales groups are as follows[19]: ecology (23), refugees (21), agriculture (17), national values (16), scandals (13), historical events (12), denials (12), human rights (11), science (10), tourism (8), media (7), infrastructure (7), influential on non-political issues" (6), violence (2), and culture (1).

Rationales provide useful insight for check-worthy claim detection task. Our main observations are as follows. Firstly, topic of a claim is an important factor to make it check-worthy. In particular, economics, politics and social issues are the most common topics in check-worthy claims. Secondly, people might have different rationales to consider a claim as check-worthy. For instance, three annotators considered the following tweet as check-worthy but provided rationales at different groups (namely, health, politics, and social issues): "The copays for medical examination and drugs will increase by 60% and 70%, respectively, on Monday". While the claim is clearly about health, other annotators considered it check-worthy for its impact on politics and social life. Thirdly, annotators frequently expressed negative outcomes or problems in their rationales (e.g., "ecological problems", "infraction of rules on a important topic", "illegal action", "problems of refugees", and "injustice"). This suggests that claims about negative issues are more likely to be check-worthy than claims about positive issues. Fourthly, many rationales we collected are about possible impacts of claims such as "it might affect people's political stance", "it can change the perception towards an influential person in the history", and "people might make an important decision based on this claim". Therefore, impact of claims should be also considered in detecting check-worthy claims. Lastly, rationales like "national values" and "about a topic important for the society" suggest that check-worthiness of some claims are not universal but they are check-worthy for just a particular nation/country. Therefore, check-worthy claim detection models should also consider national issues of each country.

[19]The numbers in parentheses represent how many times a rationale from the respective group appears in TrClaim-19.

7 Baseline Results

In this section, we provide performance results of four models on our dataset in order to provide reference baselines for future studies. We first randomly selected 635 claims (out of 765) gathered from Teyit and DP, and used 1900 tweets retrieved using these selected claims for training. The remaining 387 tweets are used for testing. The baseline models we use are as follows.

- **M-BERT:** As the best performing models in CLEF 2020 Check That! Lab (Barrón-Cedeno et al., 2020) used variants of BERT model (Devlin et al., 2019), we fine tune multiligual cased version of BERT (i.e., M-BERT) using traning data.

- **BERTurk:** Pires et al. (2019) report that M-BERT's performance might decrease for the languages with different word orders in sentences. Turkish language has also different word ordering than English. Therefore, we fine tune BERTurk (Schweter, 2020) which is a monolingual BERT model pre-trained using only Turkish texts.

- **Logistic Regression with Bag-of-Words (LR-BOW):** We train a logistic regression model with bag-of-words features. In particular, we first apply the following preprocessing techniques: case folding, removing non-alphabetic characters, eliminating stopwords with NLTK[20], and stemming[21]. We tokenize using NLTK and eventually, have 6157 distinct words to be used as the bag-of-words features.

- **Support Vector Machines with Bag-of-Words (SVM-BOW):** We use the same bag-of-words features and train an SVM model with the training data.

We use Scikit toolkit[22] for both SVM-BOW and LR-BOW models with default parameters. We adjust all models in two settings: binary and multiclass classification. In binary classification we use aggregated labels (i.e., "check-worthy" and "not check-worthy") based on majority voting for

[20] www.nltk.org/
[21] https://snowballstem.org
[22] https://scikit-learn.org

Table 5: Evaluation results for baseline models on TrClaim-19. The best result for each metric is **boldfaced**.

Model	AP	P@1	P@5	P@10	P@30	R-P	nDCG
M-BERT	**.3825**	**1.0000**	**.8000**	**.4000**	.3000	.3510	.8708
BERTurk	.3687	0.0000	.6000	.4000	.3000	**.3709**	**.8895**
BOW-LR	.3609	**1.0000**	.2000	.2000	.2667	.3245	.8815
BOW-SVM	.3716	**1.0000**	.2000	.2000	**.3333**	.3444	.8372

each tweet. In multiclass classification, we use total number of check-worthy labels for each tweet, yielding a quaternary (4-point) scale of labels (i.e., 0, 1, 2, and 3). We rank tweets based on their check-worthiness using our baseline models. We report average precision (AP), R-Precision (R-P), and precision with different cutoff values (P@k) for binary classification. For multiclass classification, we report nDCG scores. **Table 5** shows the evaluation results for the baseline models.

Based on AP, P@1, P@5, and P@10, M-BERT yields the highest scores. BERTurk model outperforms others based on R-P and nDCG metrics. As expected, BERT based models outperform LR-BOW and SVM-BOW models in most of the cases. However, their scores are close in many cases, and SVM-BOW and LR-BOW slightly outperform M-BERT and BERTurk models in various cases. This might be because of the size of the training dataset. Future studies might explore weak supervision or cross lingual transfer learning to better fine tune BERT models.

8 Conclusion

In this work, we introduce TrClaim-19, which is the first annotated dataset for Turkish check-worthy claims. We first crawled 225M Turkish tweets in 2019 by tracking keywords about important events happened in Turkey. In order to select the tweets to be annotated, we first crawled 765 claims fact-checked by two Turkish fact-checking websites. Then we used these claims as search queries to select 3 tweets to be annotated for each claim. In total, we collected annotations for 2287 Turkish tweets. In addition to check-worthy annotations, we also collected rationales behind their judgments. Furthermore, we provide performance results of baseline models on TrClaim-19 for future studies.

In our analysis of TrClaim-19, we have the following observations. Firstly, annotators have low agreement on check-worthiness of claims, suggesting that we need graded judgments, instead of binary judgments. Secondly, in many cases, non-experts disagree with experts on check-worthiness of claims even though they knew that they have been fact-checked by experts. Thirdly, annotators might have different rationales for labeling a claim as check-worthy. Fourthly, the rationales we collect suggest that topics of claims are the main factors affecting their check-worthiness. In particular, claims about economics, politics, and society are likely to be check-worthy. Lastly, the rationales suggest that claims about negative events are more likely to be check-worthy than claims about positive events.

In the future, we plan to extend to our work covering other languages. Besides being useful data resources, they will also allow us to compare rationales across different languages and cultures. In addition, we plan to apply a think-aloud methodology to collect rationales, allowing us to better understand the thought process of annotators. Furthermore, collecting rationales for not-check-worthy claims might be useful to understand disagreement between annotators.

References

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness. In *CLEF (Working Notes)*.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer.

Federica Cherubini and Lucas Graves. 2016. The rise of fact-checking sites in europe. *Reuters Institute for the Study of Journalism, University of Oxford*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. 2018. Measuring the reach of "fake news" and online disinformation in europe. *Reuters institute factsheet*.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez i Villodre, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *RANLP*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *CIKM '15*.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10:1945–1948.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30.

Cédric Lespagnol, Josiane Mothe, and Md Zia Ullah. 2019. Information nutritional label and word embedding to estimate information check-worthiness. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 941–944. ACM.

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. *HCOMP*, 16:139–148.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, pages 355–370. Springer.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.