Shared-task campaigns such as NIST TREC select documents to judge by pooling rankings from many participant systems. Therefore, the quality of the test collection greatly depends on the number of participants and the quality of submitted runs. In this work, we investigate i) how the number of participants, coupled with other factors, affects the quality of a test collection; and ii) whether the quality of a test collection can be inferred prior to collecting relevance judgments. Experiments on six TREC collections demonstrate that the required number of participants to construct a high-quality test collection varies significantly across different test collections due to a variety of factors. Furthermore, results suggest that the quality of test collections can be predicted.
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.
The high cost of constructing test collections led many researchers to develop intelligent document selection methods to find relevant documents with fewer judgments than the standard pooling method requires. In this paper, we conduct a comprehensive set of experiments to evaluate six bandit-based document selection methods, in terms of evaluation reliability, fairness, and reusability of the resultant test collections. In our experiments, the best performing method varies across test collections, showing the importance of using diverse test collections for an accurate performance analysis. Our experiments with six test collections also show that Move-To-Front is the most robust method among the ones we investigate.
While sentiment analysis is a popular research area, most of the research has been conducted for English and the number of studies for Turkish are rather limited. Limited resources for Turkish natural language processing (NLP) is one of the major challenges for Turkish NLP research. In order to overcome these limitations, we propose two approaches for Turkish sentiment analysis: 1) fine tuning multilingual model of BERT 2) using main model of BERT after machine translation of Turkish texts into English. We conducted experiments on Turkish movie and hotel review datasets where each review is labeled either positive or negative. Our methods achieve high accuracy scores such that in the movie dataset, our BERT models outperform existing methods.
In this paper, we propose an automatic text summarization model for Turkish news articles using machine learning models. Our proposed model uses sentence position, speech expression, presence of named entities and statements, term frequency and title similarity as features. We construct and share a new dataset for Turkish text summarization. In our experiments, we show that all our features we use have a positive impact on the performance of the system. In addition, we show that our model outperforms the latent semantic analysis based baseline method.
When collecting item ratings from human judges, it can be difficult to measure and enforce data quality due to task subjectivity and lack of

transparency into how judges make each rating decision. To address this, we investigate asking judges to provide a specific form of rationale supporting each rating decision. We evaluate this approach on an information retrieval task in which human judges rate the relevance of Web pages for different search topics. Cost-benefit analysis over 10,000 judgments collected on Amazon's Mechanical Turk suggests a win-win. Firstly, rationales yield a multitude of benefits: more reliable judgments, greater transparency for evaluating both human raters and their judgments, reduced need for expert gold, the opportunity for dual-supervision from ratings and rationales, and added value from the rationales themselves. Secondly, once experienced in the task, crowd workers provide rationales with almost no increase in task completion time. Consequently, we can realize the above benefits with minimal additional cost.

To create a new IR test collection at low cost, it is valuable to carefully select which documents merit human relevance judgments.Shared task campaigns such as NIST TREC pool document rankingsfrom many participating systems (and often interactive runs as well)in order to identify the most likely relevant documents for humanjudging. However, if one's primary goal is merely to build a testcollection, it would be useful to be able to do so without needingto run an entire shared task. Toward this end, we investigate mul-tiple active learning strategies which, without reliance on systemrankings: 1) select which documents human assessors should judge;and 2) automatically classify the relevance of additional unjudgeddocuments. To assess our approach, we report experiments on fiveTREC collections with varying scarcity of relevant documents. Wereport labeling accuracy achieved, as well as rank correlation whenevaluating participant systems based upon these labels vs. full pooljudgments. Results show the effectiveness of our approach, and wefurther analyze how varying relevance scarcity across collectionsimpacts our findings. To support reproducibility and follow-onwork, we have shared our code online.

Public procurement constitutes an important part of economical activities. In order to effectively use public resources, increasing competition among firms participating in public procurement is essential. In this work, we investigate the impact of content information on the number of bidders in public procurement. We explore 6 different groups of features including n-grams, named entities, language of notices, country of the authority, description length, and CPV codes. In our experiments, we show that our proposed models outperform all baselines. In particular, k-nearest neighbor model with n-grams achieves the best prediction accuracy. Our model can be used by public procurement officials to automatically examine procurement notices and detect the ones causing low competition. Besides, participating firms can use our model to predict potential competition they will face, and make better decisions accordingly.

Social media platforms such as Twitter provide an incredibly efficient way to communicate with people. While these platforms have many benefits, they can also be used for deceiving people, spreading misinformation, manipulation, and harassment. Social bots are usually employed for these kind of activities to artificially increase the amount of a particular post. To mitigate the effects of social bots, many bot detection systems are developed. However, the evaluation of these methods are challenging due to lack limited available datasets and the variety of bots people

might develop. In this work, we investigate vulnerabilities of state-of-the-art Botometer social bot detection system by creating our own bot scenarios instead of using public datasets. In our experiments, we show that Botometer is not able to detect our social bots, showing that we need more enhanced bot detection models and test collections to better evaluate systems' performances.
The scarcity of Arabic test collections has long hindered information retrieval (IR) research over the Arabic Web. In this work, we present ArTest, the first large-scale test collection designed for the evaluation of ad-hoc search over the Arabic Web. ArTest uses ArabicWeb16, a collection of around 150M Arabic Web pages as the document collection, and includes 50 topics, 10,529 relevance judgments, and (more importantly) a rationale behind each judgment. To our knowledge, this is also the first IR test collection that includes rationales of primary assessors (ie, topic developers) for their relevance judgments, exhibiting a useful resource for understanding the relevance phenomena. Finally, ArTest is made publicly-available for the research community.
On June 24, 2018, Turkey conducted a highly consequential election in which the Turkish people elected their president and parliament in the first election under a new presidential system. During the election period, the Turkish people extensively shared their political opinions on Twitter. One aspect of polarization among the electorate was support for or opposition to the reelection of Recep Tayyip Erdoğan. In this paper, we present an unsupervised method for target-specific stance detection in a polarized setting, specifically Turkish politics, achieving 90% precision in identifying user stances, while maintaining more than 80% recall. The method involves representing users in an embedding space using Google's Convolutional Neural Network (CNN) based multilingual universal sentence encoder. The representations are then projected onto a lower dimensional space in a manner that reflects similarities and are consequently clustered. We show the effectiveness of our method in properly clustering users of divergent groups across multiple targets that include political figures, different groups, and parties. We perform our analysis on a large dataset of 108M Turkish election-related tweets along with the timeline tweets of 168k Turkish users, who authored 213M tweets. Given the resultant user stances, we are able to observe correlations between topics and compute topic polarization.
The massive amount of misinformation spreading on the Internet on a daily basis has enormous negative impacts on societies. Therefore, we need automated systems helping fact-checkers in the combat against misinformation. In this paper, we propose a model prioritizing the claims based on their check-worthiness. We use BERT model with additional features including domain-specific controversial topics, word embeddings, and others. In our experiments, we show that our proposed model outperforms all state-of-the-art models in both test collections of CLEF Check That! Lab in 2018 and 2019. We also conduct a qualitative analysis to shed light-detecting check-worthy claims. We suggest requesting rationales behind judgments are needed to understand subjective nature of the task and problematic labels.
Misinformation has many negative consequences on our daily life. While spread of misinformation is very fast, investigating veracity of claims is slow. Therefore, we urgently need systems helping human fact checkers in the combat against misinformation. In this paper, we present our participation in check-worthiness tasks (ie, Task 1 and Task 5) of CLEF-

2020 Check That! Lab. For English Task 1, we use logistic regression with fined-tuned BERT predictions, POS tags, controversial topics and a hand-crafted word list as features. For English Task 5, we again use logistic regression with fined-tuned BERT predictions and word embeddings as features. For Arabic Task 1, we use a hybrid approach of fined-tuned BERT model with the model used for English Task 5. For the Arabic task, we use AraBert as our Bert model. In the official evaluation of primary submissions, our primary models a) ranked 3rd in Arabic Task 1 based on P@ 30 and shared the 1st rank with another group based on P@ 5, b) ranked 5th in English Task 1 based on average precision and shared the 1st rank with five other groups based on reciprocal rank, P@ 1, P@ 3 and P@ 5 metrics, and c) ranked 3rd in Task 5 based on average precision.