”Truth”-seeker Dataset Exploration and Analysis

Justin Johnson

Abstract

The proliferation of misinformation on social media platforms has created a pressing need for automated systems to assess the truthfulness of online content. This paper critiques and recreates the exploration of Dadkhah et al. 2023 in the use of machine learning models, specifically the DistilBERT model (Sanh et al., 2019), to predict the ”truthfulness” or ”consensus” of tweets in relation to a given statement. I introduce a methodology that incorporates both the tweet content and associated statement into a single sentence, which is then used to train the model. My experiments focus on training the model to predict the agreement between a tweet and a statement based on majority consensus labels derived from human annotators. I discuss the implications of my findings, suggest potential improvements, and propose directions for future research in the domain of automatic truthfulness detection in online content.

1 Introduction

In recent years, the need for automated systems to assess the truthfulness of online content has become increasingly urgent. With the rise of social media, misinformation, and disinformation have posed significant challenges for both individuals and institutions. Identifying whether a statement made in a post is accurate or false is a critical task in mitigating the spread of misinformation. This paper critiques the exploration done for this problem (Dadkhah et al., 2023) by utilizing machine learning models, particularly the DistilBERT model (Sanh et al., 2019), to predict the ”truthfulness” of a tweet based on its content.

In this work, I focus on recreating the experiment of training a model to determine the agreement between a tweet and a given statement, which I prefer to refer to as ”consensus”. The dataset used in this study comprises tweets and associated statements, along with labels indicating whether the tweet agrees with the statement, derived from human consensus. I explore various preprocessing strategies to prepare the data and evaluate the effectiveness of these strategies through model accuracy.

2 Background & Related Works

The work presented in this paper is situated within the broader context of research on fact-checking and sentiment analysis. Below, I summarize the relevant literature and its connection to my study.

2.1 Fact-Checking Practices

Lee et al. (2023) conducted an extensive data-driven examination of fact-checking practices across four prominent fact-checkers: Snopes, PolitiFact, Logically, and the Australian Associated Press FactCheck. Their study analyzed over 22,000 fact-checking articles, focusing on agreement and variation in verdicts, and noted that major events, such as the COVID-19 pandemic and presidential elections, significantly impacted fact-checking frequency. This work relates closely to my research as it highlights the intricacies of aggregating consensus from multiple fact-checkers, a concept central to the ”consensus” (or ”ground truth” as the Truthseeker Dataset authors suggested) labels used in this paper.

2.2 Sentiment Analysis in Text

The field of sentiment analysis has been a critical component of natural language processing research. Varghese and Jayasree (2013) provide a literature survey emphasizing the role of sentiment analysis in mining opinions from unstructured textual data, particularly from customer reviews on e-commerce platforms. They underscore the challenges in identifying emotions and opinions within text data, which directly aligns with my exploration of sentiment analysis as a tool to determine the agreement between tweets and statements.

2.3 Deep Learning for Sentiment Analysis

Building on sentiment analysis, Tyagi et al. (2020) explore the application of deep learning methods, particularly CNN-LSTM architectures with pretrained embeddings, to classify sentiments in large-scale unstructured data such as social media content.

2.4 Sentiment Analysis on Twitter

Kharde et al. (2016) offer a survey of sentiment analysis methodologies specifically targeting Twitter data. They review machine learning and lexicon-based approaches, discussing their applicability to highly unstructured and heterogeneous opinion streams on social platforms.

2.5 Relevance to This Work

The studies discussed above collectively inform the methodologies and tools applied in my research. While fact-checking studies like Lee et al. (2023) inspire the conceptual framing of consensus labels, sentiment analysis research, particularly from Varghese and Jayasree (2013) and Kharde et al. (2016), demonstrates the feasibility and challenges of analyzing sentiment and opinions in text. Moreover, Tyagi et al. (2020)'s focus on deep learning corroborates the suitability of transformer-based models, such as DistilBERT, for handling these tasks.

3 Truthseeker Dataset

The TruthSeeker dataset paper presents a large-scale, crowd-sourced dataset for the purpose of detecting real vs. fake content in social media, especially focused on tweets. The dataset is designed to help in the development of models for fact-checking and truth detection by associating statements with a set of tweets discussing them. These tweets are labeled based on whether they agree with or refute the truth of the statement they discuss.

3.1 Terminology

The TruthSeeker Dataset paper (Dadkhah et al., 2023) introduces an interesting and novel dataset for assessing the truthfulness of statements based on social media content. This exploration is interesting, because the harms of misinformation are real, so investigating and creating systems to combat it are critical. However, some aspects of its terminology could be critiqued for clarity, and consistency, to ensure we are truly achieving those aims:

- **Ambiguity** in the Use of ”Truth”: One of the main issues with the terminology in the paper is the reliance on the term ”truth” to describe the dataset's classification goal. The concept of ”truth” is inherently complex and multifaceted, particularly in the context of social media, where claims are often subjective, context-dependent, and influenced by individual beliefs. While the paper aims to classify statements as ”true” or ”false,” this binary simplification overlooks the nuances of truthfulness.

- **Over-reliance** on ”Ground Truth”: The concept of ”ground truth” is frequently used throughout the paper to refer to the verified accuracy of statements, usually provided by professional fact-checkers, who often don't even agree with each other (Lee et al., 2023). While this is a standard approach in the machine learning and information retrieval fields, the authors do not sufficiently explore the limitations of using fact-checker labels as the ”ground truth.” Fact-checking itself can be subjective, particularly when different organizations use varying methodologies or interpret evidence differently. By equating ”ground truth” with ”absolute truth,” the paper risks oversimplifying the complexities inherent in fact-checking, and may give the false impression that a single ”truth” exists for every statement. A more cautious approach would involve acknowledging the limitations of the fact-checking process and presenting ”ground truth” as a best estimate.

- ”Agreement” as a Label: The use of ”agreement” as a label for statements in the TruthSeeker Dataset also presents challenges. The authors rely on crowd-sourced labels for ”agreement” or ”disagreement” with a statement (to be clear, this nomenclature in this particular instance isn't problematic, it is a conflated term which is also used with the ”ground truth” value to determine truthfulness), which may conflate subjective perception with objective truth. The labels from crowd-sourcing platforms often reflect the opinions or biases of the participants, rather than a definitive measure of factual accuracy. While crowd-sourcing can be a valuable method for gathering diverse perspectives, it introduces variability and uncertainty. The term ”agreement” might be better defined as ”consensus” or ”alignment” with fact-checker judgments, which would better distinguish between subjective opinion and objective factual assessment.

Clearer definitions of key terms, a more nuanced discussion of truth and agreement, and a careful consideration of the limitations inherent in the dataset would improve the paper's impact and readability, ensuring that readers can better understand and apply the dataset in the broader context of truth verification and social media analysis.

3.2 Data Preprocessing Recap

The TruthSeeker dataset was created using 700 real and 700 fake news articles from PolitiFact. For each article, 2-5 manually generated keywords were used to gather tweets via the Twitter API.

Automated keyword generation methods such as PKE, RAKE, and YAKE proved ineffective, either returning too few or too many results. Thus, manual keyword generation was preferred, resulting in a dataset of 186,000 tweets (133 tweets per article on average).

3.3 Crowdsourcing and Labeling

Tweets were labeled using Amazon Mechanical Turk (MTurk), with tasks assigned to Master Turkers to assess each tweet's agreement with the source statement. The agreement levels were labeled as: *True*, *False*, *Mostly True*, and *Mostly False*. Each task was completed by three independent Turkers for accuracy.

3.4 Preprocessing

After data collection, rows with a "NO MAJORITY" or "Unrelated" label were removed. A "ground truth" column was created based on the majority answer to determine the truthfulness of each tweet. The final dataset contains 150,000 unique tweets and 1,400 statements, balanced evenly between true and false statements.

Statement (T/F)	Majority Answer
T	Agree
T	Disagree
T	Mostly Agree
T	Mostly Disagree
F	Agree
F	Disagree
F	Mostly Agree
F	Mostly Disagree

Table 1: 4-Label Conversion Truth Table.

Statement (T/F)	Majority Answer
T	Agree
T	Disagree
F	Agree
F	Disagree

Table 2: 2-Label Conversion Truth Table.

3.5 Additions & Changes

Initially, I focused on training the model using the "consensus" labels, which were derived by matching the majority answer with the "ground truth", as shown in Table 1 and Table 2. However, after further analysis, I decided to shift my approach and use the majority answers directly as the target labels. This approach yielded significantly better results.

In addition to the methodology outlined by Dadkhah et al. (2023), I implemented a data splitting strategy that ensured unique statements in each set, guaranteeing that no statements from the validation or test sets appeared in the training set. This helped prevent data leakage and ensured a cleaner evaluation process.

To enhance the model's training, I also introduced a new feature column that combined the "ground truth" value, the statement, and the tweet into a single sentence. This new feature allowed the model to better capture the relationship between the tweet and the statement, improving its ability to assess agreement.

Finally, I converted this enriched DataFrame into a 'datasets' Dataset object for compatibility with the 'transformers' model.

4 Evaluation

My intent for exploring this dataset (Dadkhah et al., 2023) was to try and recreate their experiment to the best of my ability. After doing so, I would then focus on improvements and other explorations. Due to the time constraints I have for researching and submitting this as an assignment, I have not yet explored everything I would like to explore. The following details what I have covered so far.

4.1 Experiment 1: Consensus

The goal of this first attempt was to train the "Agreement" or "truthfulness" label. In my code, I referred to this label as "consensus" instead, in an effort to explore new terminology. Throughout these experiments, I focused primarily on accuracy as my metric for improvement and evaluation. I did not explore other alternatives, which could be a worthwhile direction for future work.

Task & Procedure Using information from Tables 1 and 2, I extracted the respective majority answers and compared them against the "ground truth" as shown in those tables to derive a consensus. I then used DistilBERT (Sanh et al., 2019) to fine-tune and train the model on these consensus labels.

It is important to note that, unlike all other experiments in this paper, for this particular experiment I did not blend statements and tweets. Instead, I focused solely on the tweets and trained the model to determine "Truth" (or as I prefer to call it, "consensus") based on the tweet content.

Results This approach resulted in a relatively poor performance, achieving only about a 30% accuracy. This outcome can likely be attributed to the fact that the tweets alone, without the full context of the associated statements, may lack sufficient information for accurate agreement prediction. Tweets often contain ambiguous or incomplete information, which likely led to the lower accuracy in classifying them as agreeing or disagreeing with the statements.

I did not attempt to rerun this particular experiment with the proper preprocessing steps I later developed and used. If I have the time and opportunity to do so, I will replace this comment with a new experiment section detailing that endeavor.

4.2 Experiment 2: 4-class Classification

The primary goal of this endeavor was to train on a 4-way classification task. The targets were values of "True", "Mostly True", "False", "Mostly False".

Task & Procedure This experiment was an improvement on the prior experiment, in that I started preprocessing the data "properly". I included the statement, tweet, and "ground truth" values in a single sentence, which was then fed to my DistilBERT (Sanh et al., 2019) model. I then trained with the "majority_answer" as my label, in order to determine how well I could predict if a tweet agreed with a given statement.

I also played around with parameters to try and *squeeze* out some extra performance. For example, even with a relatively small learning rate (1e-5), I was getting oscillating loss (high to low very rapidly). I had to reduce this to 1e-8 to get the chart seen in Figure 1.

But even so, I reached a stable accuracy value after a single epoch, and didn't see much improvement upon changing any of these other parameters.

Results This experiment resulted in similar results achieved by Dadkhah et al. (2023) 's team, achieving nearly 49% accuracy on the test and validation sets. I believe that this had a hard time making predictions, mostly because there is no real difference between "True" and "Mostly True". Those are extremely subjective ideas, especially considering how this team crowd-sourced the labels.

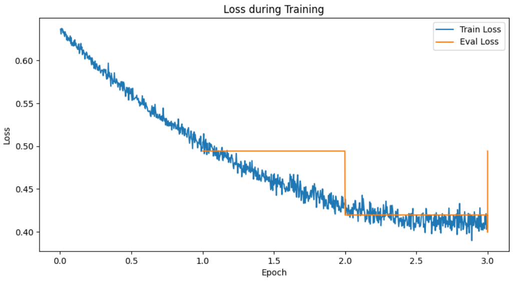

Figure 1: 4 Label Class Evaluation Loss

4.3 Experiment 3: 2-class Classification

This was a two-way classification task, between "True" and "False"

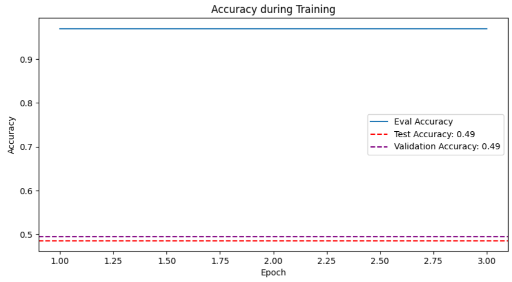

Figure 2: 4 Label Class Evaluation Accuracy

Task & Procedure This experiment closely shadows the previous one, except the focus was binary classification. For multi-class classification in the prior experiment, I swapped to use CategoricalLoss instead of BinaryCrossEntropyLoss. Otherwise the preprocessing and analysis remained fairly similar.

Results This experiment also resulted in similiar outcomes as Dadkhah et al. (2023). I achieved about 96% accuracy.

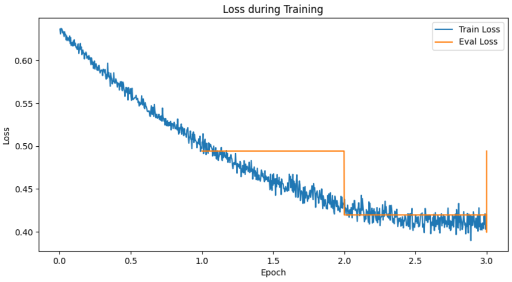

Figure 3: 2 Label Class Evaluation Loss

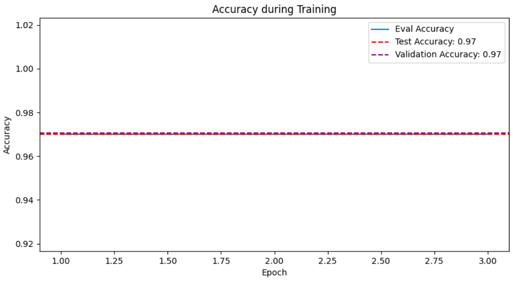

Figure 4: 2 Label Class Evaluation Accuracy

This also had small variation when I attempted to adjust parameters. After a single epoch, the model didn't learn much more it would seem. But it was able to properly predict the majority answer much better than the prior experiment.

Both this and the prior experiment could be used in a prediction pipeline to do what my 1st experiment attempted to do. Since the agreement score is fairly well learned, I could then just match that with the provided ”ground truth” statement (supposing that it is supplied) to determine ”truth”. Such a process should result in similar accuracies, since this is a matter of simple computation and not AI learning.

4.4 Experiment 4: Sentiment Analysis

In this experiment, I explored the role of sentiment analysis in determining the agreement between tweets and statements, with a focus on attempting to ascertain truth with only the contents of a tweet. I used two methods: DistilBERT (Sanh et al., 2019) for deep learning-based sentiment analysis, and VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014) for lexicon-based sentiment extraction. The goal was to investigate whether sentiment, in the form of polarity scores, could be leveraged to predict how well a tweet agreed with a given statement.

Task & Procedure I found that using VADER's (Hutto and Gilbert, 2014) sentiment analysis output helped me quantify the sentiment of each tweet, which I then compared to the sentiment of the corresponding statement. However, I also realized that while sentiment scores from VADER and DistilBERT (Sanh et al., 2019) provided useful information about the emotional tone of tweets, they did not directly correlate with the agreement between tweets and statements in terms of ”truth” or consensus.

Class	Precision	Recall	F1-Score	Support
False	0.74	0.61	0.67	41
True	0.69	0.80	0.74	45
Accuracy	0.71 (86)			
Macro avg	0.71	0.70	0.70	86
Weighted avg	0.71	0.71	0.71	86

Table 3: Sentiment Analysis Classification Report

Results Precision measures the proportion of positive predictions among all positive predictions. In this case, the model performed slightly better when predicting ”False” sentiment (74% precision) compared to ”True” sentiment (69% precision). This suggests that the model is more reliable in predicting negative sentiment than positive sentiment. However, the difference is small, indicating a balanced performance across both classes.

Recall measures the proportion of actual positives correctly identified by the model. The model performed better in identifying positive sentiment (”True”) with a recall of 80%, compared to negative sentiment (”False”) with a recall of 61%. This indicates that the model is more sensitive to positive sentiment, but misses more negative sentiment instances.

The F1-score is a harmonic mean of precision and recall. The model achieved a slightly better F1-score for ”True” sentiment (0.74) than for ”False” sentiment (0.67). This suggests that the model is more balanced in its performance when predicting positive sentiment.

The overall accuracy of the model is 71%, meaning that 71% of the instances in the test set were correctly classified. This indicates a decent level of overall performance, though there is still room for improvement.

This analysis pointed to the fact that sentiment alone—whether positive or negative—was not enough to predict whether a tweet would agree with a statement. It was a valuable exploratory step, but I concluded that sentiment needed to be combined with other features (Kharde et al., 2016), such as factual accuracy or opinion extraction (Varghese and Jayasree, 2013), to improve the model's performance in predicting agreement. Future work could explore how these features could be integrated to better assess the relationship between tweet sentiment and statement agreement.

5 Implications

The results and methods discussed in this paper carry several important implications for both the field and the broader community. These should be carefully considered in future research and application:

- **Potential for Bias:** The reliance on crowd-sourced data to generate majority answers and labels, as described in this study, could unintentionally introduce biases based on the demographics or perspectives of those providing the data. If not appropriately mitigated, such biases may distort the accuracy and fairness of the consensus reached, leading to misclassifications that affect decision-making processes in real-world applications.

- **Misuse in Sensitive Areas:** The ability to predict agreement between statements and tweets could be applied in contexts where sensitive or controversial topics are discussed. Misuse of such tools to enforce specific viewpoints or manipulate public opinion could be harmful. There is a risk of algorithmic amplification of polarized opinions, especially in politically charged or socially sensitive areas.

- **Impact on Public Discourse:** The results of this study may also influence public discourse by automating content moderation or sentiment analysis. While such tools have the potential to streamline information processing, they could inadvertently silence marginalized voices or prevent meaningful debate if not carefully calibrated. This may contribute to echo chambers or the suppression of alternative viewpoints.

- **Vulnerabilities in System Design:** The approach used in this study may be vulnerable to adversarial attacks, such as manipulation of the input data (e.g., tweets or statements) to achieve a desired outcome. For instance, adversarial actors could craft specific tweets that align with a false consensus, thus manipulating the model's predictions. Proper safeguards and testing for robustness against such attacks are essential.

- **Ethical Considerations:** As AI-based systems are increasingly integrated into decision-making processes, it is critical to ensure that the systems are transparent and accountable. This study showcases the need for clear ethical guidelines regarding the use of AI models for social or political applications. Transparency in how consensus is determined and ensuring that no harmful consequences arise from mislabeling or misunderstanding public opinion are essential to the responsible development of such systems.

6 Future Research & Work

One potential avenue for future research involves exploring alternative methods for determining "consensus." In this study, Hutto and Gilbert (2014)'s sentiment analysis tool could be utilized to determine agreement between statements and tweets, providing an automated approach to calculating consensus. This contrasts with the crowdsourced method used in Dadkhah et al. (2023), which relied on human annotators via Amazon Mechanical Turk. While Dadkhah et al. (2023) presents an interesting and widely applicable method, the reliability of crowdsourced labels can be questioned due to inconsistencies across annotators (Lee et al., 2023).

In contrast, Hutto and Gilbert (2014) claims to outperform humans in sentiment classification, particularly on social media data, and while sentiment is not synonymous with agreeableness, it is worth investigating whether VADER could provide a more scalable and efficient solution for this task. A comparison between crowdsourced consensus and VADER's output could reveal insights into the strengths and limitations of both approaches, offering directions for improving automated consensus detection in future work.

Additional future research directions include:

- **Exploring Other Sentiment Tools:** Besides VADER, other sentiment analysis models like BERT-based classifiers or RoBERTa (Liu et al., 2021) could be tested to see if they can improve consensus prediction, particularly when trained on specific tweet data and statements.

- **Evaluation of Model Robustness:** It would be important to evaluate the robustness of the models trained for consensus detection. This includes testing for adversarial attacks, where misleading tweets could be crafted to manipulate the predicted consensus, potentially undermining the validity of automated systems.

- **Multimodal Data Fusion:** Future research could consider incorporating both tweet text and associated metadata (e.g., author sentiment, number of likes/shares, and time posted) as additional input features for consensus prediction. Combining multiple data sources may improve model performance and better reflect the real-world factors that influence agreement.

- **Expanding Consensus Metrics:** While this research primarily focused on binary classification (i.e., agreement or disagreement), a more nuanced approach could be explored. For example, determining the strength or intensity of agreement, not just a simple binary response, could lead to more useful and insightful predictions in various contexts, such as political discourse or product reviews.

- **Cross-Domain Applications:** Exploring the application of consensus prediction methods in other domains, such as fake news detection, scientific consensus, or public health, could broaden the impact of this research. Each domain might require a slightly different approach, and adapting consensus prediction models could provide valuable contributions in various fields.

These avenues would help further refine the methodologies for automated consensus detection, and could have practical applications in real-time data processing, sentiment tracking, and improving public discourse analysis.

Conclusion

The primary goals of this research were to recreate the Truthseeker experiment, explore sentiment analysis as a tool for determining consensus, and investigate potential improvements to their methods in terms of methodology, terminology, and process. In recreating the Truthseeker experiment, I focused on using sentiment analysis techniques, specifically VADER and DistilBERT, to classify tweets and statements according to their agreement with a given statement.

While the experiments showed promising results, it became clear that further refinement is needed, particularly in integrating sentiment analysis models with consensus detection and improving the accuracy of the predictions. One key area for future work is exploring how sentiment models, like VADER, could outperform or complement human-sourced consensus labels.

These ideas open the door to improving the methodology behind consensus detection and sentiment analysis, with the potential for future advancements in areas such as fake news detection, social media monitoring, and automated opinion analysis. Future work should further refine the model, explore new avenues for improving accuracy, and continue to examine the role of sentiment analysis in determining agreement in textual data.

References

S. Dadkhah, X. Zhang, A. G. Weismann, A. Firouzi, and A. A. Ghorbani. 2023. The largest social media ground-truth dataset for real/fake content: Truthseeker. *IEEE Transactions on Computational Social Systems*, 99:1–15.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee. 2023. “fact-checking” fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Vishu Tyagi, Ashwini Kumar, and Sanjoy Das. 2020. Sentiment analysis on twitter data using deep learning approach. In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)*, pages 187–190. IEEE.

Raisa Varghese and M Jayasree. 2013. A survey on sentiment analysis and opinion mining. *International journal of Research in engineering and technology*, 2(11):312–317.