DESIGNING AND DEBIASING BINARY CLASSIFIERS FOR IRONY AND SATIRE DETECTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ASLI UMAY ÖZTÜRK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2024

Approval of the thesis:

DESIGNING AND DEBIASING BINARY CLASSIFIERS FOR IRONY AND SATIRE DETECTION

submitted by **ASLI UMAY ÖZTÜRK** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, Graduate School of **Natural and Applied Sciences** __________

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** __________

Prof. Dr. Pınar Karagöz
Supervisor, **Computer Engineering, METU** __________

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU __________

Prof. Dr. Pınar Karagöz
Computer Engineering, METU __________

Assoc. Prof. Dr. Burkay Genç
Computer Engineering, Hacettepe University __________

Date:05.09.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Aslı Umay Öztürk

Signature :

ABSTRACT

DESIGNING AND DEBIASING BINARY CLASSIFIERS FOR IRONY AND SATIRE DETECTION

Öztürk, Aslı Umay
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Pınar Karagöz

September 2024, 96 pages

In the age of social media, detecting ironic and satirical text automatically is a challenging task that is important for fighting misinformation online. Even though there are compelling datasets and research conducted in other languages, the literature lacks any large datasets and comprehensive studies conducted in Turkish. This work aims to fill that gap by first curating two datasets for irony and satire detection, and uses curated datasets to explore binary classification pipelines for irony and satire detection tasks with traditional supervised learning methods such as SVM (Support Vector Machine) and large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers). Furthermore, this work discusses the possible biased nature of the curated datasets by stylistic analysis, and possible inherited bias of the trained models by using model explainability methods and comparing the results with human annotations. Finally, a pipeline is proposed for debiasing and improving model generalisability by using synthetic data generation with LLMs.

Keywords: Irony detection, Sentiment analysis, Natural language processing, Debi-

asing, Large language models, Text generation

ÖZ

İRONİ VE SATİR TESPİTİ İÇİN İKİLİ SINIFLANDIRMA MODELLERİNİN TASARLANMASI VE ÖNYARGIDAN ARINDIRILMASI

Öztürk, Aslı Umay
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Pınar Karagöz

Eylül 2024 , 96 sayfa

Sosyal medya çağında, ironi ve mizahi metinleri otomatik olarak tespit etmek, çevrimiçi yanlış bilgilere karşı mücadele için önem arz etmektedir. Diğer diller için oluşturulmuş kapsamlı veri setleri ve yapılmış detaylı araştırmalar bulunmasına rağmen, Türkçede büyük bir veri seti ve kapsamlı bir çalışma literatürdeki önemli eksiklerden biridir. Bu çalışma, ironi ve mizah tespiti için iki veri seti hazırlayarak bu boşluğu doldurmayı amaçlamaktadır. Hazırlanan veri setlerini kullanarak, ironi ve mizah tespiti problemleri için SVM (Destek Vektör Makineleri) gibi geleneksel denetimli öğrenme yöntemleri ve BERT (Çift Yönlü Kodlayıcı Temsilleri) gibi büyük dil modelleri (LLM) ile ikili sınıflandırma modelleri tasarlanmıştır. Ayrıca bu çalışma, metinlerde stil analizi yöntemleriyle oluşturulan veri setlerinin taraflı olup olmadıklarını ve model açıklanabilirlik yöntemlerinden alınan sonuçların insan açıklamaları ile karşılaştırılmasıyla da modellerin taraflı ya da önyargılı olup olmadıklarını incelemektedir. Son olarak, LLM'ler ile sentetik veri üretimi yapılarak modelin önyargısını giderme ve genellenelebilirliğini artırma için bir metod önerilmektedir.

Anahtar Kelimeler: İroni tespiti, Duygu analizi, Doğal dil işleme, Önyargıdan arındırma, Büyük dil modelleri, Metin üretme

Tijo'ya

ACKNOWLEDGMENTS

A master thesis is supposedly the work of one person, but I am only able to complete this work thanks to the help of Onur, D.Eren, Birkan, Helin, M.Ozan, Özgür, Eda and Yiğit.

Of course, I could never be here without the support of my family, I will forever be grateful for their unconditional support.

I would also like to thank Pınar Karagöz and Cem Bozşahin for being my mentors, the Data Mining Research Group people for all the hard work we did together, and everyone who helped me during my time here at METU CENG.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
DT	Decision Tree
kNN	k-Nearest Neighbours
LIME	Local Interpretable Model-Agnostic
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long-Short Term Memory
ML	Machine Learning
NB	Naive Bayes
NELA	News Landscape
NLP	Natural Language Processing
RF	Random Forest
RQ	Research Question
SHAP	Shapley Additive Exlanations Explanations
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

CHAPTER 1

INTRODUCTION

Thanks to the increased accessibility of technology and the internet in the modern world, information is now widely available in many forms to be consumed by the average user. Unfortunately, this implies that any unsuspecting user can easily interact with misinformation [1].

The generation and spread of misinformation can occur accidentally or on purpose. On-purpose examples are more straightforward: a user or several users may maliciously generate and spread fake stories for several reasons including propaganda [1, 2]. On the other hand, accidental misinformation can be more nuanced. For instance, one user can accidentally take an ironic or satirical piece of content to be true and interact with it as if it is the truth. With a cascading effect, the satirical or ironic content the misinformation originated from could even be taken as the truth by several people.

Automated labelling of misinformation is a functionality that has been in demand by social media platforms in the last decade, with several platforms trying to show automated warnings next to shared posts, with several studies discussing their effectiveness [3]. We believe detecting ironic and satirical content is important to differentiate them from maliciously generated misinformation.

From a different point of view, irony and satire detection are also important to extract accurate information about the authors of a text. This is crucial for opinion mining tasks that are utilised in e-commerce and other product/service-based areas. For instance, a negative review of a user may be taken as positive by a sentiment analysis model if it fails to capture the irony in the review text, creating a biased analysis of the reviewed product. This analysis may affect business or social decisions, and on a large scale, may result in monetary or reputation loss.

Unfortunately, the problem of detecting irony and satire is not a straightforward task even for humans, and there are different challenging factors when it comes to Natural Language Processing (NLP) methods. One challenge is the availability of high-quality labelled data with high volume. Since data is limited for most languages and the nature of both irony and satire inherently encodes the style of the writer, the success of the proposed models may be misleading and the results may not always be generalisable.

1.1 Motivation and Problem Definition

Keeping the importance of irony and satire for accurate opinion representation in mind, the motivation of this thesis is as follows: this work aims to design Turkish language centric irony and satire detection models, analyse their generalisability performance and biases, and then propose a debiasing approach to improve them.

Subproblems of this thesis work are:

- Irony detection: binary classification of a text to belong into either "ironic" or "non-ironic" label
- Satire detection: similarly, binary classification of a text to belong into either "satiric" or "legitimate" label
- Understanding the bias of a dataset: analysis of a dataset to see if it is inherently skewed from a statistical perspective, causing the models that are trained with it to be biased
- Debiasing of a model: reducing the bias of a model which makes decisions that are skewed in a specific manner (can be biased towards a label, or more attentive to a specific style of text, etc.)

1.2 Proposed Methods and Models

This work proposes irony and satire detection pipelines utilising both traditional and large language model based methods, and demonstrates their performance on newly curated datasets for Turkish. Furthermore, this work explores the bias of the designed models and curated datasets to propose a debiasing pipeline for improving model performance and generalisability.

1.3 Contributions and Novelties

Contributions of this work are as follows:

- Curation of the IronyTR Dataset, the largest irony dataset available for Turkish informal texts
- Curation of the Turkish Satirical News Dataset, the largest satirical news dataset for Turkish with extended human annotations
- Design and comparison of different methods and feature extraction approaches for irony detection in informal Turkish texts
- Design and comparison of different methods and feature extraction approaches for satire detection in Turkish news texts
- Analysis and discussion of the bias of irony and satire detection models and datasets for Turkish
- Proposal of a debiasing pipeline for irony and satire detection models using large language models, prompt engineering, and text generation

1.4 The Outline of the Thesis

This thesis is organized around three separate but related studies, where each study is discussed in their respective chapters. Chapter 2 goes over the literature review done for each study in separate sections with some overlaps. Chapter 3 focuses on the design of the irony detection models for Turkish short-form texts. Chapter 4 discusses the design of the satire detection models for Turkish news texts. Chapter 5 presents the biased findings of the models from the previous chapters and discusses debiasing approaches. Finally, this work is concluded in Chapter 6 with a comparative discussion of Chapters 3, 4, and 5.

CHAPTER 2

RELATED WORK

2.1 Introduction

This work combines the literature on binary text classification, specifically the identification of ironic, satiric, or non-factual texts. On top of these, the literature on debiasing models and datasets is explored extensively. This chapter summarises the literature reviewed during the research of this thesis, divided into 4 sections, namely, *Definition of Irony, Sarcasm, and Satire*, *Binary Classification of Irony, Sarcasm, and Satire*, *Satirical News, Fake News, Misinformation*, and *Explaining Model Decisions and Debiasing*.

2.2 Definition of Irony, Sarcasm, and Satire

There are no standardised and universally agreed definitions for irony and sarcasm, specifically in Natural Language Processing (NLP) literature. This causes different studies to use these terms interchangeably, or even for different concepts. Several literary works and NLP research exist that discuss a separation between sarcasm and irony [4], but most research uses them interchangeably [5]. On the other hand, satire is generally studied separately, even though sometimes it is used in relation to both irony and sarcasm [6].

There seems to be a gap in the literature for accurate definitions of these terms, which may be explored in collaboration with linguists to improve the understanding and detection of such phenomena in human-created content. The rest of this section refers to the concepts with the words the authors of the respective research used to define their tasks.

Even though the separation between these concepts is not studied, subcategories of irony are discussed in several studies [7, 8]. This categorisation is explored to improve model performance, introducing the need for context for some subcategories of irony. On the other hand, a similar need for social context is also discussed in satire studies [6], making the subcategories of irony questionable to some extent.

2.3 Binary Classification of Irony, Sarcasm, and Satire

Starting from the mid-2010s, the interest in several subproblems of sentiment analysis has increased. In irony and sarcasm detection, a similar spike occurred, with different challenges and benchmark datasets being published.

One such example is the third task of SemEval-2018[1], which was focused on the irony detection problem in English tweets [7]. The task utilized the English part of the dataset collected and annotated by Van Hee et al. [8], which was labeled as *ironic*, *possibly ironic*, and *non-ironic*. In the same paper, Van Hee et al. also proposed a Dutch dataset for irony detection. Per a similar trend, IroSvA Task[2] in IberLEF 2019 published the Spanish data for irony detection with different variants [9]. There are also independent studies with relatively smaller datasets in Chinese [10], Turkish [11], French and Arabic [12]. Similar to irony detection, the sarcasm detection problem is mainly explored on English datasets [5], with relatively small studies in other languages [13].

On the side of detection models, the literature includes studies that mainly used English datasets and supervised learning methods [5, 14, 15, 16, 17].

One of the earlier studies is the study of Buschmeier et al. [15], where they combine different feature sets and classification methods to compare their performance on irony detection. They use a dataset of product reviews with star ratings, and utilise these ratings on top of several lexical and syntactic features. The supervised learning methods they use include Support Vector Machine (SVM), Random Forest (RF),

[1] https://competitions.codalab.org/competitions/17468
[2] https://sites.google.com/view/iberlef-2019/home?pli=1

Logistic Regression (LR), Decision Tree (DT) and Naive Bayes (NB) classifiers.

Another one of the earlier studies is the study of Barbieri et al. [5] where they tackle the problem in six different domains/categories and compare the performance of a tree-based classifier. They also propose new lexical features such as frequency (gap between rare and common words) or intensity (intensity of adverbs and adjectives) that aim to capture sarcasm.

Van Hee et al., the curators of the dataset shared in the third task of SemEval-2018, approach the problem of irony detection in a more detailed manner in their 2016 study [17]. They first classify tweets into ironic and non-ironic categories, furthermore, they define three subcategories: situational irony, irony by means of polar clash, and other verbal irony. They use an SVM-based classifier and a feature set with basic lexical, syntactic, semantic and sentimental features. On the other hand, Pamungkas and Patti [16] using a smaller feature set, additionally exploit the sentiment analysis of emojis in the text for more information.

Baloglu et al. [14] also compare several supervised machine learning algorithms based on K-Nearest Neighbours (k-NN), DT, and several others on a similar set of features. In another study, Ahmed et al. [18] exploit graph representations of sentences to extract similarity-based features.

Many of the studies mention that situational irony is harder to detect. Addressing this issue, Carvalho et al. [6] model different rhetorical devices used to give an ironic meaning to text with lexical and syntactic features, and define a measurement of predictability degree for a situation. Using these features, they aim to model out-of-domain contrast. They conclude that this contrast measure improves the performance of the model they trained and tested on a Portuguese data set of farcical news headlines. Another claim in the literature considers the negative effect of sarcasm in the performance of sentiment analysis tasks, where Tan et al. [19] combine both problems as a multi-task learning approach. Their work shows promising results.

The literature also includes neural network based approaches. For instance, Wu et al. [20] propose an LSTM (Long-Short Term Memory) based detection model where they also utilise additional features such as Part of Speech (PoS) tags, sentiment features, and sentence embeddings. On the other hand, Zhang et al. [21] proposes an approach to this problem with transfer learning from the sentiment analysis task. Their results show that transferred sentiment increases the ability of the model to detect context incongruity.

On the last couple years, newer studies are focusing more and more on multimodal approaches, and LLM-based models. In their study, Tomas et al. [22] explore the irony detection performance of transformer-based models with both textual and visual inputs. On the other hand, Lin et al. [23] combines transformer and LLM based models with prompt engineering to improve the irony detection performance, specifically focusing on different features of the text.

There exist several studies conducted on languages other than English [6, 24, 12, 10], but only a handful of studies exist for Turkish, mainly done as preliminary work or on small datasets. For instance, Dulger [25] studies the irony classification on a balanced Turkish data set of 144 instances, using SVM, k-NN, NB, RF classifiers as well as LR and Multilayer Perceptron, working with a limited set of features.

In their 2017 study, Taslioglu and Karagoz [26] work on irony classification on a larger and balanced Turkish data set of 194 instances. Their study also includes polarity score based sentimental features, similar to the approach of Van Hee et al. [17].

Another study conducted on Turkish, which is a preliminary version [27] of the work in this thesis, compares the performances of SVM, NB and LSTM based classifiers and BERT *(Bidirectional Encoder Representations from Transformers)* [28] on a balanced Turkish data set of 220 instances.

2.4 Satirical News, Fake News, Misinformation

In the late 2010s, similar to irony and sarcasm detection, a spike of interest also occurred for satirical news, fake news, and misinformation detection on social media. There exist several curated datasets for fake news detection problem in English, each focusing on different aspects, topics, or following a different annotation scheme [29, 30, 31]. There are also several studies focusing on satirical news in English [32, 33, 34, 35, 36]. Combining both, there are also studies discussing the difference between satirical and fake news with curated datasets [37, 38].

In the literature, the problem of satirical news detection can be explored in two different categories: detecting satirical or ironic short-text content or detecting fake and deceptive content. These problems also differ on a structural level. In most cases, fake or deceptive content exists in a network-based environment, where there is a network of people interacting with the content. For satirical or ironic content, the interaction information is not widely available, or in some cases, non-existent. Hence, while fake news detection studies rely on network-based information or fact-checking websites, satire detection studies heavily rely on linguistic analyses.

On the fake content detection side, there are several studies conducted both in English [33, 39], and in other languages such as Portuguese [6]. On top of the approaches from the sarcastic, satirical, and ironic short text classification, these studies also include additional contextual information from fact-checking websites to verify the reliability of the news articles. Several studies try to analyze user interaction graphs and the credibility of the author to understand the legitimacy of the shared content [40].

Fake news detection focusing on Turkish landscape is also studied by several researchers. Koru & Uluyol[41] and Taskin et al.[42] focus on tweets for identifying fake content. On the other hand Mertoğlu & Genç[43] focus on news articles collected from different resources, proposing an automated system to fact-check content. However, their dataset is not publicly available.

Satirical news is an understudied topic in Turkish. To our knowledge, there exists only one work, which is by Onan and Toçoğlu [44]. They tackle the satire detection problem by using an ensemble of classifiers.

2.5 Explaining Model Decisions and Debiasing

With the rise of complex black box models, such as deep neural networks and transformers, interpreting and explaining the decisions of models have become an important task, and resulted in the creation of research fields explainability of artificial intelligence (AI) and interpretability of machine learning (ML). There exist widely used, model-agnostic methods for explainability such as LIME (Local Interpretable Model-Agnostic Explanations) [45] and SHAP (SHapley Additive exPlanations) [46], as well as studies for specific models and tasks [47, 48]. Explanations of model decisions are used to improve the performance and fairness of the model, as well as reduce the model bias.

The bias of a model may also depend on the stylistic bias of a dataset. The research of Horne et. al.[49, 50] aims to explore the stylistic difference between news articles, and they implement NELA (News Landscape) features library[3] library. NELA features library was created for news veracity detection, but is also utilised more generally in other text forms. The library extracts hand-crafted, text-based features in six categories including the style, complexity, and bias of the article.

Qian et al. [51] further focus on dataset bias, and propose a framework for debiasing using counterfactual inference. Their show that their approach improves the effectiveness, generalisability, and fairness of the classifier.

In another study, Schlicht et al. [52]utilise conversational LLMs to reduce textual bias in news articles. Their findings show that even though they are compelling in some cases, they tend to leave out vital and contextual information during the debiasing process.

[3] https://github.com/BenjaminDHorne/NELAFeatures

CHAPTER 3

IRONY DETECTION IN SHORT-FORM TURKISH TEXTS

This chapter mostly covers one of the already published works of the thesis author [11]. Only the parts the author is the main contributor is included in this chapter.

3.1 Introduction

With the latest advancements in technology, humans started to use the Internet for most of their daily tasks. As a result, there exists almost an abundance of online textual content with varying styles, lengths, and contexts. This abundance makes the automated processing of text more important, because understanding user opinion from a large set of texts is only possible via automation.

Irony detection, which is a subproblem of sentiment analysis is important to conclude correct results from automated opinion mining processes. Opinions extracted are important for a wide set of applications, including service and product improvements, e-commerce, or public relations. There exists a rich set of methods for sentiment analysis, but these methods are not reliable when irony is present, especially in Turkish texts. Oxford Dictionary defines irony as *the expression of one's meaning by using language that normally signifies the opposite, typically for humorous or emphatic effect*[1]. This definition makes it visible that irony is not easy to detect: the signification of the opposition is not explicit in most cases, and sentiment analysis methods do not have the "common sense" that us humans have.

For instance, consider the following example:

[1] https://www.lexico.com/en/definition/irony

COVID-19 toplantısı, COVID-19 önlemleri kapsamında iptal edildi.
COVID-19 meeting is cancelled due to COVID-19 precautions. (transl.)

The irony of this text comes from the situation, and can be hard to capture without any context. On the other hand, some forms of irony can be captured by textual features, for instance, the following sentence is a good example of irony by means of a polar clash between the polarity of its words:

Sabrımı denememeni *çok* ***seviyorum****!*
I just **love** *when you* ***test my patience!*** (transl.)

Such polar clashes can be explored using feature engineering and utilising sentiment analysis methods. Hence, feature engineering has the potential to improve the performance of irony detection models.

This part of the thesis explores the problem of irony detection in Turkish short-form online texts, that are by their nature, mostly informal (such as tweets, microblog entries, comments, etc). The goal is to analyse and improve the performance of the methods that are studied in English and other languages to Turkish, on a new and comprehensive data set, in order to see their performance in Turkish. The structure of this section is as follows:

- Section 3.2 explains the data collection process of the *IronyTR: Extended Turkish Social Media Dataset for Irony Detection*, which was curated within the scope of the research for this thesis.

- Section 3.3 explains the methods that are utilised in the experiments, preprocessing, and feature extraction pipelines.

- Section 3.4 describes the experiments and results.

- Section 3.5 discusses the results only in the scope of this part.

3.2 IronyTR: Extended Turkish Social Media Dataset for Irony Detection

During the time of this study, the only openly available Turkish dataset for the irony detection task was the previous version of the IronyTR dataset with a much smaller instance count, which we have also curated in an earlier study [27]. On the other hand, there were relatively big datasets for English and Spanish that were curated and published for workshops and challenges [8, 24]. To work on this gap in the literature, we followed a similar path to the aforementioned studies to collect data, and curated the *IronyTR: Extended Turkish Social Media Dataset for Irony Detection*.

3.2.1 Data Collection

Data is mostly sourced from Twitter (now known as X)[2]. and Eksisozluk[3]. All instances are authentic user entries, and only minor edits are applied to remove offending words or shorten the data instances.

For Twitter, the API was used to collect batches of data, with keywords such as "#ironi", "#sarcasm" and "#irony", as well as new tweets from current trending topics. The collected data was then inspected to remove irrelevant entries. This phase does not include annotation for labelling, which is performed separately.

For Eksisozluk, popular topics of the week and all-time popular topics were scanned manually. Again, this phase does not include annotation for labelling, which is labeled separately.

3.2.2 Annotation

For the annotation, 7 native Turkish speakers are asked to label the data as "ironic", "non-ironic", and "unsure". If 5 or more annotators have agreed on the "ironic" label, the data is labelled as "ironic". Similarly, if 5 or more annotators have agreed on the "non-ironic" label, the data is labelled as "non-ironic". The rest of the collected data was discarded.

[2] https://x.com
[3] https://eksisozluk.com

The remaining dataset is reduced to include a balanced set of 300 ironic and 300 non-ironic instances. The dataset is openly available on GitHub page[4].

3.3 Methods

This section describes the methods utilised for the experiments, as well as the preprocessing and feature extraction pipelines. All methods mentioned here use the *IronyTR* dataset.

3.3.1 Preprocessing and Feature Extraction

Before extracting features from instances, firstly, each instance is preprocessed by tokenisation of the words, punctuation marks, and emojis/emoticons. All letters are converted to lowercase. An example preprocessing pipeline works as follows:

"Sınava geç kaldım, aferin bana!" is transformed into the following tokens: *"sınav geç kalmak , aferin ben !"* where tokens are *"sınav", "geç", "kalmak", ",", "aferin", "ben", "!"*.

The feature extraction pipeline follows the preprocessing task. A set of syntactic and lexical features, which are used in several studies [17, 24, 25, 26] are extracted using the preprocessed data:

- **Word Count:** A float value indicating the ratio of words to all tokens
- **Interjections:** A binary value indicating if any interjection words ("bravo", "oley (transl. hurray)" etc.) exist
- **Boosters:** A binary value indicating if any booster words ("asla (transl. never)", "mutlaka (transl. of course") etc.) exist
- **Repetition:** A binary value indicating if there are any repeated tokens in the sentence

[4] https://github.com/teghub/Turkish-Irony-Dataset

- **Capitalization:** A binary value indicating if there are any capitalized words in the sentence (extracted before converting every letter to lowercase)
- **Emoji/Emoticons:** Two features, one being a binary value indicating if any emojis/emoticons exist, the other being a float value indicating the ratio of the count of emoji/emoticons to all token count
- **Exclamation Marks:** Two features, one being a binary value indicating if an exclamation mark exists, the other being a float value indicating the ratio of count to all punctuation mark count
- **Question Marks:** Similar to exclamation mark features, a binary value indicating if a question mark exists, the other being a float value indicating the ratio of count to all punctuation mark count
- **Ellipsis:** Similar to exclamation mark features, a binary and a float value for the ellipsis
- **Quotation Marks:** Similar to exclamation mark features, a binary and a float value for quotation marks
- **Bracketed Exclamation Marks:** Similar to exclamation mark features, a binary and a float value for bracketed exclamation marks
- **Bracketed Question Marks:** Similar to exclamation mark features, a binary and a float value for bracketed question marks
- **All Punctuation Marks:** A float value indicating the ratio of punctuation mark count to all token count
- **Bag of Words:** A vector of the size of the whole corpus, where the count of each token in the sentence is shown with a normalized float value.

It should be noted that for some tokens such as punctuation marks and emoji/emoticons, both a float and a binary feature are extracted because the existence of a token and the ratio of a token to other tokens may have a different impact on the representation of the data.

On top of these features, one important approach that is more focused on the definition of irony is to extract polarity-based features. By its definition, an ironic statement includes opposing concepts, which may be inferred from the sentiment or polarity score of its words. This idea is built on the sentiment analysis studies that widely utilise the polarity scores of words [18, 25, 26, 17, 53].

Since there was no publicly available polarity score look-up library for Turkish during the time of this study, following the approach of other studies, existing English libraries were translated manually [54]. Using *SenticNet*[5] the words in the dataset are manually translated to create a look-up table. This look-up table includes a score between -1.0 and 1.0 for each word in our collection. A sample polarity scoring for the sentence *"Sınava geç kaldım, aferin bana!"* is shown in Figure 3.1.

Figure 3.1: Polarity scores of the words in a sentence

The following features are extracted using the look-up table:

- **Average Polarity:** Three float values representing the sum of the positive, negative, and all polarity score values to sentimental token ratio
- **Minimum Polarity:** A float value indicating the minimum polarity score existing in the sentence
- **Maximum Polarity:** A float value indicating the maximum polarity score existing in the sentence
- **Maximum Polarity Difference:** A float value indicating the difference of minimum and maximum polarity, scaled to be between 0 and 1
- **Positive and Negative Sum Difference:** A float value indicating the difference of positive polarity score sum and negative polarity score sum, scaled to be between 0 and 1

[5] https://www.sentic.net/downloads/

- **Polarity Contrast:** A binary value indicating the existence of both positive and negative polarity scores in the sentence

This study also integrates a graph representation based method to explore hidden relationships within the text. The core idea here is to create class graphs for both labels and somehow capture the similarity of an instance to both of these graphs to extract a new feature point.

Utilising the method described in the study by Ahmed et al. [18], a sentence graph is created for each instance in the dataset with a vicinity window of 3. The graph representation has a directed edge from each token to its following two tokens. However, it should be noted that only the words and emoji/emoticon tokens are used as vertices in the graphs. As an example, the graph constructed for the sentence *"Sınava geç kaldım, aferin bana!"* is shown in Figure 3.2. For simplicity, the graph is created with a vicinity window of 2.

Following the creation of sentence graphs, class graphs are created by taking the union of the sentence graphs for each instance belonging to that class.

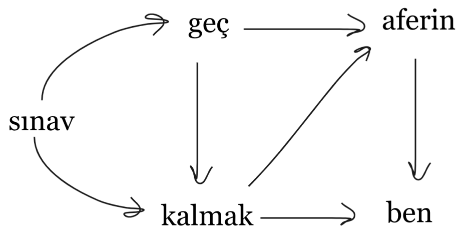


Figure 3.2: Sample sentence graph constructed for the sentence "Sınava geç kaldım, aferin bana!" with vicinity window of 2.

Using the graphs, the following features are extracted:

- **Containment Similarity Scores:** Two float values each one representing the containment similarity score of a sentence graph to ironic and non-ironic class graphs. Containment similarity is calculated as given in Equation 3.1.

$$\frac{|S \cap C|}{min(|S|, |C|)} \tag{3.1}$$

where S is the sentence graph, C is the class graph, and $|graph|$ representation stands for the size of the graph. In this definition, size of a graph and the size of the intersection of two graphs are calculated with respect to the number of edges.

A summary of the preprocessing and feature extraction pipeline can be seen in Figure 3.3.

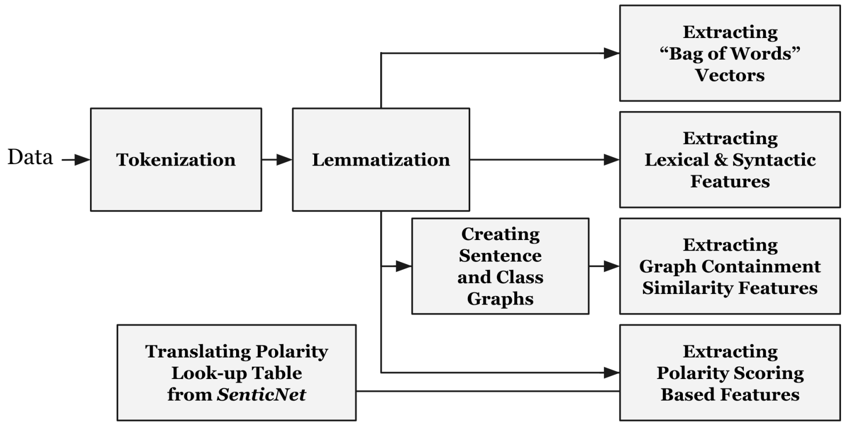


Figure 3.3: Preprocessing and feature extraction pipeline

3.3.2 Learning Methods

Coverage of this thesis includes three traditional supervised learning methods and one large language model based neural method. The research that this chapter is based on also includes several deep neural network methods that are omitted here [11].

Learning methods utilised in the experiments can be described as two separate items:

- **Genetic Optimization with *TPOT*[6] for Traditional Supervised Learning Methods:** *TPOT*, an automated machine learning tool that searches a space to find the optimal model is used to decide whether to choose SVM (Support Vector Machine), Multinomial Naïve-Bayes (NB), or Decision Tree (DT) for a specific pipeline. It should be noted that the other optimization functionalities

[6] https://epistasislab.github.io/tpot/

of TPOT, such as feature selection, are not used in this study. TPOT is only used to optimize model parameters.

- **BERT (Bidirectional Encoder Representations from Transformers):** *BERT*[7] is a masked large language model (LLM) developed by Google Research. This study utilises *BERT Base Multilingual Cased* pre-trained model, which is fine-tuned by the researcher for binary classification of textual data.

3.4 Experiments and Results

The experiments are constructed to include different pipelines, each utilising a subset of extracted features or raw data. Table 3.1 summarises the feature/method combinations for each pipeline. While traditional supervised learning methods use these features to represent the data, BERT directly takes the tokenized text to create its own representation, hence it is not included in the table.

3.4.1 Traditional Supervised Learning Methods with TPOT

The performance of 5 main pipelines optimized using TPOT to decide between traditional supervised learning methods Multinomial NB, SVM and DT were analysed. The pipelines mainly differ by the feature subset used and are described in detail below. For each pipeline, the performance results of the optimized classifiers are compared in Table 3.2.

- **"Baseline" Pipeline** The baseline method uses Bag of Words vectors since they contain the minimal information that can be extracted from a sentence.

- **"Basic" Pipeline:** For the basic features pipeline, as seen in Table 3.1, on top of the BoW vectors, lexical and syntactic features are included in the feature set.

[7] https://github.com/google-research/bert

- **"Polarity" Pipeline:** Since polarity scores can contain information to facilitate irony detection, another pipeline is created by adding polarity-based features to the "Basic" pipeline (as shown in Table 3.1).
- **"Graph" Pipeline:** Graph similarities can also contain important or hidden information. To utilise graph similarities, another pipeline is created by adding the containment similarity score features to the features used in the "Basic" pipeline (as shown in Table 3.1).
- **"Pol-Gra" Pipeline:** Finally, both graph-based and polarity-based features are added on top of the features used in the "Basic" pipeline to create the "Pol-Gra" pipeline.

3.4.2 BERT Pipeline

For this study, a 12-layered BERT model is created using an open-source implementation[8] with added changes for weight freeze implementation. Hyperparameter settings (5 epochs, 0.00004 learning rate, 16 batch size) are kept the same when trying different weight freeze settings. All results are evaluated by 10-fold cross-validation and are shown in Table 3.3, where *Layers* indicate the last number of layers that are not frozen, i.e., layers where the parameters are trained.

3.4.3 Evaluation Metrics

The experiments are conducted under 10-fold cross-validation and accuracy, precision, recall and F1-score metrics are reported. These metrics are calculated by using their common definitions involving the number of *true positive*, *true negative*, *false positive* and *false negative* instances.

3.4.4 Traditional Pipeline Results

TPOT is used to choose the best-performing classifier for each pipeline, and the selection is done by comparing the F1-score. The highest score by column for each

[8] https://github.com/ThilinaRajapakse/pytorch-transformers-classification

pipeline, as well as the selected pipelines, are written in bold in Table 3.2. The table shows that incrementally adding more features results in a trend of increasing performance for all classification methods. Comparing the polarity-based and graph-based features separately, a higher performance improvement is seen for polarity-based features, with a bigger spike in SVM. But when selecting the best performing for all pipelines using F1-score, Multinomial Naïve Bayes classifier is preferable.

3.4.5 BERT Pipeline Results

To select the best-performing BERT settings, performance metrics for different training layers are evaluated and reported in Table 3.3. The results show that 6-layer freeze has a better performance, hence it is chosen to be compared with other models in Table 3.4.

3.5 Discussion

Table 3.4 shows the comparative performance of the best-performing settings for each pipeline. The main findings of this study are as follows:

- As expected, the chosen "Baseline" pipeline with only Bag of Words (BoW) has a low performance.
- As hypothesised, incremental addition of features improved the performance of traditional learning methods.
- On the other hand, BERT pipeline shows the best performance in terms of accuracy, precision and F1-score, without any feature engineering.
- In general, traditional pipelines provide higher recall results than the BERT pipeline.

In conclusion, this part of the thesis analyses the performance of traditional supervised learning methods as well as BERT for irony detection in Turkish informal texts. Firstly, a dataset of 600 instances is created, and the analysis is conducted on this newly curated dataset, which is a larger Turkish dataset created for this purpose. The effects of feature engineering using polarity score and graph-based features on recognizing irony analysed. It is observed that polarity score and graph-based features improve the performance of traditional classifiers. On the other hand, BERT outperforms all pipelines and gives more promising results with no feature engineering.

Table 3.1: Features Used in Pipelines

Features	Baseline	Basic	Polarity	Graph	Pol-Gra
Bag of Words	x	x	x	x	x
Word Count		x	x	x	x
Emojis/Emoticons		x	x	x	x
Interjections		x	x	x	x
Boosters		x	x	x	x
Repetition		x	x	x	x
Capitalization		x	x	x	x
Exclamation mark		x	x	x	x
Question mark		x	x	x	x
Ellipsis		x	x	x	x
Quotation mark		x	x	x	x
Bracketed Excl.		x	x	x	x
Bracketed Qs.		x	x	x	x
All Punctuation		x	x	x	x
Avg Polarity			x		x
Min Polarity			x		x
Max Polarity			x		x
Max Polarity Diff			x		x
Pos-Neg Polarity Diff			x		x
Polarity Contrast			x		x
Containment Similarity				x	x

Table 3.2: Comparison of Traditional Methods

Pipeline/Method	Accuracy	Precision	Recall	F1-score
Baseline				
SVM	48.17%	29.64%	**56.33%**	38.84%
MultinomialNB	48.17%	39.54%	56.32%	**46.46%**
Decision Tree	**52.17%**	**52.19%**	35.17%	42.02%
Basic				
SVM	53.50%	60.01%	66.34%	63.02%
MultinomialNB	55.33%	58.83%	**68.44%**	**63.27%**
Decision Tree	**56.67%**	**64.58%**	39.74%	49.20%
Polarity				
SVM	**63.33%**	**64.57%**	58.22%	61.23%
MultinomialNB	55.83%	58.23%	**74.35%**	**65.31%**
Decision Tree	53.50%	58.64%	56.56%	57.58%
Graph				
SVM	53.50%	60.01%	66.34%	63.02%
MultinomialNB	**58.17%**	**61.40%**	**67.90%**	**64.49%**
Decision Tree	54.83%	58.86%	62.43%	60.59%
Pol-Gra				
SVM	**63.33%**	**65.53%**	57.10%	61.03%
MultinomialNB	55.67%	58.11%	**76.12%**	**65.91%**
Decision Tree	55.50%	56.00%	50.40%	53.05%

Table 3.3: Comparison of BERT with Different Trained Layers

Layers	Accuracy	Precision	Recall	F1-score
11	68.06%	68.92%	63.67%	66.19%
7	64.83%	66.46%	60.17%	63.16%
6	**69.00%**	**71.34%**	**65.75%**	**68.43%**
5	66.50%	67.23%	61.64%	64.31%
1	63.50%	65.21%	59.76%	62.37%

Table 3.4: Comparison of Methods

Pipeline	Accuracy	Precision	Recall	F1-score
Baseline	48.17%	39.54%	56.32%	46.46%
Basic	55.33%	58.83%	68.44%	63.27%
Polarity	55.83%	58.23%	74.35%	65.31%
Graph	58.17%	61.40%	67.90%	64.49%
Pol-Gra	55.67%	58.11%	**76.12%**	65.91%
BERT	**69.00%**	**71.34%**	65.75%	**68.43%**

CHAPTER 4

SATIRE DETECTION IN TURKISH NEWS

This chapter is founded on the findings of two unpublished studies conducted with two co-authors, but the work described here is only the author's own. The work of other authors is omitted in order not to take credit from them, or included with disclaimers/citations.

4.1 Introduction

With the increased usage of social media, for many people, the primary source of news and information has become the shared news articles in their social media feeds. Even though this makes the information more accessible, it can also cause misinformation to spread at fast rates. It is not uncommon for regular social media users to take fake or satirical content as the truth, which is specifically problematic when it comes to news content. Satire detection, which is a subproblem of sentiment analysis, can offer a solution to this misinformation problem since automated detection of satirical content can be used to create automated warnings that inform social media users about the reliability of a piece of information.

Cambridge dictionary defines satire as *humorously criticizing people or ideas, especially to make a political point*[1]. Consider the headlines:

(1) İmamoğlu: "Seçilirsem kadınlar 8 Mart'ta çalışmayacak"
(transl.) İmamoğlu: "If I am elected, women will not work on March 8th"

[1] https://dictionary.cambridge.org/dictionary/english/satirical

(2) SpaceX'in Fırlatılması Öncesi Konuşan ABD Başkanı Trump, Göstericileri Hedef Aldı: "Ne yaparsanız yapın, Mars'a kutlu yürüyüşümüzü durduramayacaksınız..."
(transl.) POTUS Trump spoke before the SpaceX launch and targeted the protesters: "Whatever you do, you will not be able to stop our blessed march to Mars..."

When one only reads the headlines shared in short social media posts, they may not understand the content is in fact satirical. In the example headlines, headline (1) is correct information, but headline (2) is a piece of satirical news shared by the popular satirical newspaper Zaytung[2]. Here, to understand the satire in the headline (2), one needs some context and specific background information about certain political figures. And similarly, without any context, one can easily mistake headline (1) as satire.

This part of the thesis explores the problem of satire detection in Turkish news texts. A new dataset is curated for this purpose, and several satire and irony detection methods from the literature are applied to this dataset. The structure of this section is as follows:

- Section 4.2 explains the data collection process of the initial version of the *Turkish Satirical News Dataset*, and discusses the characteristics of the collected data.
- Section 4.3 explains the methods that are utilised in the experiments, preprocessing, and feature extraction pipelines.
- Section 4.4 describes the experiments and results.
- Section 4.5 discusses the results only in the scope of this part.

4.2 Turkish Satirical News Dataset

Even though there are plenty of satirical news websites in languages other than English[3], there is only a handful of datasets curated in other languages [6]. To fill this

[2] http://zaytung.com/
[3] https://w.wiki/6pR5

gap and to create a resource for satirical and fake news detection tasks in Turkish, the initial version of the *Turkish Satirical News Dataset* is curated.

4.2.1 Data Collection

As a source of satirical newspaper articles, the Turkish satirical newspaper Zaytung is used. By crawling the Zaytung website archive[4][5], 2825 satirical articles are collected with timestamp, title, body, and header image information.

As a source of legitimate newspaper articles, the 70.000+ Turkish News dataset available in Kaggle[6] is used. This dataset contains news articles from different online newspapers over a period of two months with metadata. Since this dataset is collected from different online sources, it consists of 70.000+ articles that represent a wide array of perspectives from the Turkish social environment.

To combine these two datasets into one balanced dataset with 5650 instances, 2825 articles from the 70.000+ Turkish News dataset are selected, while keeping the representative nature of the dataset.

Top 10 words from the body content of each dataset are reported in Table 4.1.

Table 4.1: Top 10 Words of two datasets

Dataset	Top 10 Words
70.000+ Turkish News	*başkan (president), bu (this), etmek (make), haber (news), maç (game/match), olmak (be), son (last), Türkiye, vermek (give), yapmak (do)*
Zaytung	*al (take), bir (one/a), demek (say), etmek (make), gelmek (come), iş (job), olarak (being), Türkiye, vermek (give), yapmak (do)*

[4] http://zaytung.com/digerleri.asp

[5] It should be noted that M.O. Alpay contributed heavily to the data collection pipeline as a co-author in a currently unpublished study.

[6] https://www.kaggle.com/datasets/suleymancan/turkishnews70000

4.3 Methods

This section describes the methods utilised for the experiments, as well as the preprocessing and feature extraction pipelines. All methods mentioned here use the initial version of the *Turkish Satirical News Dataset*.

4.3.1 Preprocessing and Feature Extraction

In this study, both the titles and the body content of news articles are used to understand if an article is satirical or not. Before any preprocessing, to be able to work on the collected data and remove any bias, the dataset is cleaned up to remove unnecessary information, such as the names of the news sources.

Two main approaches are chosen to represent each instance: representing each article by only using the title or combining both title and body content. For the article titles, the preprocessing steps are as follows:

1. Normalise the title to fix typos and convert everything to lowercase
2. Tokenise the normalized title to separate punctuation marks from the words
3. Lemmatise the tokenised title to be able to capture word similarities

Similarly, for the article body content, the preprocessing steps are as follows:

1. Normalise the body content to fix typos and convert everything to lowercase
2. Tokenise the normalized body content to separate punctuation marks from the words
3. Lemmatise the tokenised body content to be able to capture word similarities
4. Use TF-IDF scoring on the body content to extract top 10 words and discard the remaining words

For the preprocessing pipeline, Turkish NLP tool Zemberek[7] is used. After preprocessing, similar to the work described in Chapter 2 [11], following features are extracted both only for title or for title and body content combined:

- **Bag-of-Words vector:** Sums of the one-hot encodings of the words, extracted using the whole corpus
- **Basic features:** Several features that summarize the structure of the article, such as:
  - number of tokens
  - existence of '!'
  - existence of '?'
  - existence of '...'
  - existence of '"'
  - existence of booster words such as "asla (never)" or "en (most)"
  - existence of interjections such as "yaşasın (hurray)"
- **Polarity scores:** Several features that analyze the semantic difference between words using their positive or negative sentimental scores, such as:
  - minimum polarity score of a word
  - maximum polarity score of a word
  - maximum polarity score difference between pairs of words

For Bag-of-Words, well known Python machine learning library scikit-learn [55] is used. For basic and polarity features, Python scripts are implemented from scratch.

All features are used in different combinations to represent the instances in the dataset during training and testing. Details on the combinations are shown in Table 4.2.

4.3.2 Machine Learning Methods

There are several traditional supervised learning methods used in the experiments, which have been widely used in the literature.

[7] https://github.com/ahmetaa/zemberek-nlp

Table 4.2: Features Used in Pipelines

Features	BoW	Basic	Bow+Basic	Basic +Polarity	Bow+Basic+Polarity
Bag of Words	x		x		x
Word Count		x	x	x	x
Emojis/Emoticons		x	x	x	x
Interjections		x	x	x	x
Boosters		x	x	x	x
Repetition		x	x	x	x
Capitalization		x	x	x	x
Exclamation mark		x	x	x	x
Question mark		x	x	x	x
Ellipsis		x	x	x	x
Quotation mark		x	x	x	x
Bracketed Excl.		x	x	x	x
Bracketed Qs.		x	x	x	x
All Punctuation		x	x	x	x
Avg Polarity				x	x
Min Polarity				x	x
Max Polarity				x	x
Max Polarity Diff				x	x
Pos-Neg Polarity Diff				x	x
Polarity Contrast				x	x

- **Support Vector Machines (SVM):** Boser and others define SVM as "a training algorithm that maximizes the margin between the training patterns and the decision boundary" [56].

- **Decision Trees (DT):** Porgorelec and others define DT's as "a reliable and effective decision making technique that provide high classification accuracy with a simple representation of gathered knowledge" [57].

- **Multinomial Naïve Bayes (NB):** Xu and others define Multinomial NB as "a not fully Bayesian classifier that is often used as a baseline in text classification" [58].

- **k-Nearest Neighbors (kNN):** Zhang defines kNN as a classifier "used to classify unlabeled observations by assigning them to the class of the most similar labeled examples according to the distance metric" [59].

When implementing the aforementioned methods for the models, the well known Python machine learning library scikit-learn [55] is used. Configuration details of the models are described in detail in Section 4.4.

On top of the traditional methods, BERTurk [60] which is a DistilBERT-based [61] model that has been specifically trained on a Turkish text corpus [8], is used for creating the BERT pipeline. [9]

4.4 Experiments and Results

All experiments reported in this section are conducted with 10-fold cross-validation, and performances are reported with 4 metrics: accuracy, precision, recall and F1-score. Best performing models are selected using F1-score.

Experiments are designed to be repeated for two data representations: only using article titles and combining the article title with its body. For both representations, several pipelines utilising different feature combinations are created.

4.4.1 Using Only Article Titles

When building the models with using only article titles, all methods mentioned in Section 4.3 (except kNN) are trained with different feature set combinations and different hyperparameters to optimize the performance. The feature set combinations explored with only article titles are as follows:

- **Bag-of-Words:** only using the Bag-of-Words vector as feature vector
- **Basic:** only using the Basic features explained in Section 4.3 in the feature vector

[8] https://huggingface.co/dbmdz/bert-base-turkish-cased
[9] It should be noted that R.F. Çekinel contributed heavily to the implementation of the used BERT pipeline as a co-author in a currently unpublished study.

- **Basic + Bag-of-Words:** combining basic features with the Bag-of-Words vector
- **Polarity + Basic + Bag-of-Words:** adding polarity score based features to combined basic feature + Bag-of-Words vector

Methods and hyperparameters explored for training the models with only article titles are as follows:

- **SVM:**
  - Loss functions: hinge
  - C: 1e-4, 1e-2, 1e-1, 10, 100
- **DT:**
  - Criterions: Gini and Entropy
  - Maximum depth: 8, 11
- **MultinomialNB:**
  - Alpha: 0.1, 1, 10

With these settings and features, comparative results of the experiments only using the article titles are shown in Table 4.3. These results show that some of the proposed methods yield acceptable results, whereas some perform poorly while classifying satirical news content. Analyzing the performance for feature sets, it can be discussed that incremental addition of features to the feature sets increased the performance in general, with most competitive performance scores in F1 for classifying reported from the models with *Basic + BoW* and *Polarity + Basic + BoW* feature sets. The best performance scores are obtained by models that use the Decision Tree classifier. We believe that there are several possible reasons for these outcomes:

- Only using titles may result in a loss of context.
- Some supervised learning methods such as Decision Trees are susceptible to overfitting as a result of specific distinct features of our dataset.

Table 4.3: Comparison of Methods Using Only Article Titles

Features/Method	Accuracy	Precision	Recall	F1-score
Bag-of-Words (BoW)				
SVM	47.40%	48.69%	60.54%	39.84%
MultinomialNB	82.14%	75.97%	**94.36%**	83.99%
Decision Tree	**90.74%**	**92.38%**	88.86%	**90.57%**
Basic				
SVM	79.63%	74.41%	**90.36%**	81.59%
MultinomialNB	78.42%	73.77%	88.55%	80.41%
Decision Tree	**82.98%**	**86.41%**	78.31%	**82.06%**
Basic + BoW				
SVM	79.77%	74.39%	90.49%	81.59%
MultinomialNB	78.98%	72.73%	**92.90%**	81.51%
Decision Tree	**91.16%**	**92.10%**	90.09%	**91.04%**
Polarity + Basic + BoW				
SVM	83.39%	79.02%	91.03%	84.54%
MultinomialNB	81.12%	74.64%	**94.12%**	83.21%
Decision Tree	**90.93%**	**92.56%**	88.95%	**90.70%**

To overcome these issues or have a better understanding of them, a second set of experiments with a wider array of methods and hyperparameters, as well as integrating the body content of the articles, is conducted.

4.4.2 Using Both Article Title and Body Content

The pipeline for experiments that use the title and the body content for each instance is similar to the pipeline of the experiments that only use article titles. Again, all methods mentioned are trained with different feature set combinations and different hyperparameters to optimize the performance of the models. The following feature set combinations are used in the experiments:

- **Bag-of-Words:** only using the Bag-of-Words vector as feature vector
- **Basic:** only using the Basic features explained in Section 4.3 in the feature vector
- **Basic + Bag-of-Words:** combining basic features with the Bag-of-Words vector
- **Polarity + Basic:** adding polarity score based features explained in Section 4.3 to the basic feature vector
- **Polarity + Basic + Bag-of-Words:** adding polarity score based features to combined basic feature + Bag-of-Words vector

For each model, the following hyperparameter values are used in training to find the optimal performance:

- **SVM:**
  - Loss functions: hinge and squared hinge
  - C: 1e-4, 1e-2, 1e-1, 10, 100
- **DT:**
  - Criterions: Gini and Entropy
  - Maximum depth: 3, 4, 6
- **MultinomialNB:**
  - Alpha: 0.1, 1, 10
- **kNN:**
  - Weights as distance
  - Distances as Euclidean
  - Neighbors: 5, 11, 15, 19, 25, 75

With these settings and features, comparative results of the experiments with both using the article title and body content are shown in Table 4.4. It can be seen that the addition of body content and experimenting with a wider array of methods and hyperparameters increased the credibility and performance of the results. Similar to the previous experiments, the incremental addition of methods yields an overall increase in the performance for most methods. Overall, the best performing pipeline is SVM with *Polarity* + *Basic* + *BoW* feature sets combined.

Unlike the previous set of experiments, we do not see an overfitting problem with DT anymore, and we believe this is because of the change in the hyperparameter settings of the method.

Another interesting result is even though kNN fails to capture any information about the dataset when *Bow* feature set is present in the combination, when it is excluded from the combinations, we see that kNN performs better than other methods. We believe that this would make kNN a good method to use when the computational power at hand is limited.

4.4.3 Comparing the Best Performing Traditional Learning Method with BERT

Even though promising performance can be reached using traditional and simpler learning methods, reporting the performance of a transformer-based model such as BERT on this newly created dataset is important. Table 4.5 compares the scores of the BERT model trained with the best performing traditional pipeline. It can be seen that BERT can identify the satirical articles and legitimate articles almost perfectly.

This almost perfect performance seems unlikely to be scalable to real-world and real-time applications, and also raises questions about the bias of the curated dataset, creating a cascading bias for the models.

4.5 Discussion

This chapter focuses on the problem of satire detection in Turkish news text, starting with a data curation process and presenting the performance of both traditional and more state-of-the-art models. Even though the traditional models perform competitively, an almost perfect performance with BERT model is seen, which raises the following questions:

- Is the curated dataset biased since it only has a single source for satirical news?
- If there exists such a bias, how does this affect the generalisability of the models?
- Is there a feasible way for debiasing the data and/or the trained models?

We strongly believe the writing style of Zaytung News is creating a heavy bias, which can be backed by the increase in performance with the addition of body content to the representation of articles. Chapter 5 focuses on these claims and describes debiasing approaches that are explored in the scope of this thesis.

Table 4.4: Comparison of Methods Using Both Article Title and Body

Features/Method	Accuracy	Precision	Recall	F1-score
Bag-of-Words (BoW)				
SVM	**91.89 %**	91.92 %	**91.89 %**	**91.89 %**
MultinomialNB	89.47 %	89.91 %	89.47 %	89.49 %
DT	86.97 %	87.80 %	86.97 %	87.03 %
kNN	50.27 %	**99.60 %**	50.27 %	66.53 %
Basic				
SVM	79.82 %	81.72 %	79.82 %	80.01 %
MultinomialNB	79.81 %	81.02 %	79.81 %	79.91 %
DT	81.35 %	82.47 %	81.35 %	81.44 %
kNN	**85.10 %**	**85.70 %**	**85.10 %**	**85.13 %**
Basic + BoW				
SVM	86.85 %	87.08 %	86.86 %	86.87 %
MultinomialNB	**89.82 %**	90.27 %	**89.82 %**	**89.85 %**
DT	87.13 %	87.34 %	87.13 %	87.14 %
kNN	50.50 %	**99.30 %**	50.50 %	66.46 %
Polarity + Basic				
SVM	79.82 %	81.70 %	79.82 %	80.01 %
MultinomialNB	79.89 %	81.14 %	79.89 %	80.01 %
DT	82.57 %	82.70 %	82.57 %	82.57 %
kNN	**82.97 %**	**83.10 %**	**82.97 %**	**82.98 %**
Polarity + Basic + BoW				
SVM	**94.41 %**	94.45 %	**94.41 %**	**94.41 %**
MultinomialNB	90.39 %	90.77 %	90.39 %	90.41 %
DT	87.06 %	87.23 %	87.06 %	87.08 %
kNN	50.19 %	**99.67 %**	50.20 %	66.55 %

Table 4.5: Comparison of Traditional Methods and BERT

Features/Method	Accuracy	Precision	Recall	F1-score
SVM (Polarity + Basic + BoW)	94.41%	94.45%	94.41%	94.41%
BERT	99.74%	100.00%	99.33%	99.66%

CHAPTER 5

EXPLORING MODEL AND DATASET BIAS TO IMPROVE GENERALISABILITY

The research discussed in this chapter is built upon the two previous chapters. Datasets and models presented in two previous chapters are analysed for biases, and possible debiasing approaches are explored.

Some approaches and results mentioned are from unpublished works of the author with two other co-authors. The work of the co-authors is omitted or included with disclaimers and citations.

5.1 Introduction

For the last couple of years, with the rise of LLMs, an overwhelming improvement in the performance of text classification has been seen in the literature, which is also reproduced in the work discussed in Chapter 4. This raises a question of bias, which the last part of this thesis work aims to explore and improve.

5.2 Research Questions

This chapter is constructed around 3 research questions (RQ):

- **RQ1:** Can we analyse our datasets to see if they are biased stylistically?
- **RQ2:** Can we explain the decisions of our models to discuss if they are biased?

- **RQ3:** Can we design a specialised pipeline to train an unbiased model with bias-prone data?

Each RQ is discussed in its respective section with methods and results. Finally, the chapter is concluded in Section 5.8 with a discussion of the findings.

5.3 Improving *Turkish Satirical News Dataset*

Before designing the experiments, several improvements were made to the dataset:

- For the news articles with the `SATIRICAL` label taken from Zaytung, articles older than 2014 are discarded.
- For the `LEGITIMATE` label, articles resourced from an open dataset from Kaggle are discarded and replaced with scraped data from a Turkish news agency.

The final dataset includes 2202 `SATIRICAL` and 4781 `LEGITIMATE` articles. The experiments in this chapter use the improved version of the dataset described above.

5.4 RQ1: Bias of the Dataset

The first research question aims to explore the bias of the datasets. Since textual data may inherently encode the bias or the writing style of the author, and since the satirical data is taken from Zaytung, it is hypothesised that `SATIRICAL` articles are stylistically very different from `LEGITIMATE` ones.

One way to understand this is by conducting a basic statistical analysis.

5.4.1 Average Word and Sentence Count

To better understand the data instances, a primary statistical analysis is performed and reported in Table 5.2. It can be seen that data instances with `SATIRICAL` label have an average of 329 words per instance and 44 sentences per instance. On the other hand, data instances with LEGITIMATE label have an average of 313 words per instance and 43 sentences per instance. Even though the numbers are close, on average, we see that the SATIRICAL class have more words per sentence.

Table 5.1: Statistics of the *Turkish Satirical News Dataset* by labels

Statistic	SATIRICAL	LEGITIMATE
Avg. word count	329	313
Avg. sentence count	44	43

5.4.2 Top 10 Words

To have a general idea about the content of the news belonging to different labels of the dataset, top-10 terms are extracted per label by TF-IDF (Term Frequency - Inverse Document Frequency) scoring. These terms are shown in Table 5.2. It is visible that the top 10 words for the two classes do not have any words in common. This also follows the idea that the tone of the two datasets are different.

Table 5.2: Statistics of the *Turkish Satirical News Dataset* by Top 10 Words

Label	Top 10 Words
SATIRICAL	*almak (take), bir (one/a), de (also/too), etmek (make) gelmek (come), iş (work/job), olarak (being), vermek (give), türkiye (Turkiye), yapmak (do)*
LEGITIMATE	*ülke (country), yıl (year), açıklama (explanation), ifade (expression), fotoğraf (photograph), spor (sport), bölge (region), başkan (president), konu (issue), çalışmak (work)*

Literature on English proposes several other metrics and tools for stylistic analysis such as the works of Horne et al. [49, 50] implemented as the NELA features [1]

[1] https://pypi.org/project/nela-features/

library. Unfortunately, such methods rely on hand-curated dictionaries in English, and translating them to Turkish needs a linguistic background.

5.5 RQ2: Bias of the Model

The second research question aims to explore the bias of the models, specifically LLM-based models.

Training with a small set of biased data can consequently result in a biased model. Both the traditional supervised learning methods and the BERT model trained in the study in Chapter 4 show possible signs of bias. However, the same does not seem to be the case for the models in Chapter 3.

To investigate this, this part of the study aims to compare the machine and human understanding of selected instances, mainly on the *Turkish Satirical News Dataset*.

5.5.1 Understanding Human Decisions

As is, the satirical class of the *Turkish Satirical News Dataset* is labelled as satirical since it is known to be collected from a satirical online newspaper. However, it is not analysed to see what properties of the news articles make them satirical in the first place.

To act as a baseline for the explainability methods, a human annotation process is conducted for randomly selected instances in the satirical class of the dataset. The annotation process is as follows:

- The main annotator goes through the whole article body and identifies the `REAL` and `FAKE` parts.
- The `REAL` and `FAKE` markings are done according to the objective facts and events. The annotator is asked to fact-check and cite related information as needed.

- Four volunteers from different age demographics cross-check the annotations to have a higher coverage of news landscape knowledge.
- News articles with annotations that have a unified agreement are accepted, the rest is discarded.

Finally, human-annotated data consists of 40 satirical articles. Three selected annotations are shown in Figures 5.1, 5.3, and 5.5. Red text stands for the `FAKE` parts of the article, whereas the blue parts are marked as `REAL`.

5.5.2 Explaining Model Decisions

To reason about the decision of the models, we first use explainable AI and interpretable ML methods.

In a previously published study which is not fully in the scope of this thesis [48], we used LIME [45] to explain the decisions of the BERT model trained for irony detection in Chapter 3. The main finding of the study is that the BERT model makes its decision about an instance by giving importance to fewer selected words than the other models examined, but no apparent bias is detected.

To analyse the decisions of the BERT model for satirical news detection in Chapter 4, SHAP[46][2] is used. The SHAP explainability method uses Shapley values to understand the relative importance of different features for a prediction instance of a model. In other words, it assigns importance values to the features relative to each other that show their weight in the final decision.

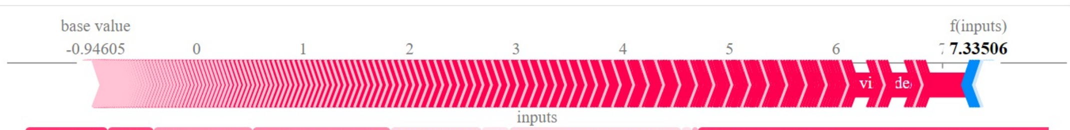
Three selected explanations are shown in Figures 5.2, 5.4, and 5.6. The red highlights in the human annotation stand for the parts of the texts that are annotated as `FAKE` and the blue highlights specify the parts that are annotated as `REAL`. Similarly, for the SHAP output, red highlighted parts are explained as the *important* parts of the texts that the classifier focuses on when identifying a data instance as `SATIRICAL`. Blue highlights in the SHAP output indicate that those parts of the texts are pulling the

[2] It should be noted that R.F. Çekinel is the main contributor to the implementation of the used SHAP pipeline as a co-author in a currently unpublished study.

label towards LEGITIMATE, and the parts that are not highlighted are not important for the decision of the model.

Yaşadığı %99.8'lik değer kaybıyla 2 gün içerisinde 64 dolar seviyesinden 0.2 dolar seviyesine gerileyen Terra Luna Coin'den kötü haberler gelmeye devam ediyor. Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından bugün bir açıklama yayınlayan Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken, "Şu son 2 günde olmaz dediğimiz ne varsa hepsi oldu. Çok üzgünüm" ifadelerine yer verdi.Durumun ekibe moral vermek için yaptığı bir toplantıda ortaya çıktığını belirten Kwon "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım. Öncesinde 'nasılsa TL'den de daha kötü durumda değilizdir' diye bakma gereği duymamıştım. Esas hata o oldu" derken, grafiği ekranda açmasıyla birlikte acı gerçeği fark ettiklerini dile getirdi. "O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz. Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten." sözleriyle Terra Labs'daki son durumu da aktaran deneyimli CEO, kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek ileriye dönük iyimser mesajlar vermeyi de ihmal etmedi.

Figure 5.1: Human annotation for news article (A)



Yaşadığı %99.8'lik değer kaybıyla 2 gün içerisinde 64 dolar seviyesinden 0.2 dolar seviyesine gerileyen Terra Luna Coin'den kötü haberler gelmeye devam ediyor. Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından bugün bir açıklama yayınlayan Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken, "Şu son 2 günde olmaz dediğimiz ne varsa hepsi oldu. Çok üzgünüm" ifadelerine yer verdi. Durumun ekibe moral vermek için yaptığı bir toplantıda ortaya çıktığını belirten Kwon "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım. Öncesinde 'nasılsa TL'den de daha kötü durumda değilizdir' diye bakma gereği duymamıştım. Esas hata o oldu" derken, grafiği ekranda açmasıyla birlikte acı gerçeği fark ettiklerini dile getirdi. "O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz. Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten." sözleriyle Terra Labs'daki son durumu da aktaran deneyimli CEO, kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek ileriye dönük iyimser mesajlar vermeyi de ihmal etmedi.

Figure 5.2: SHAP output for news article (A)

5.5.3 Comparing Human Annotations with Model Explanations

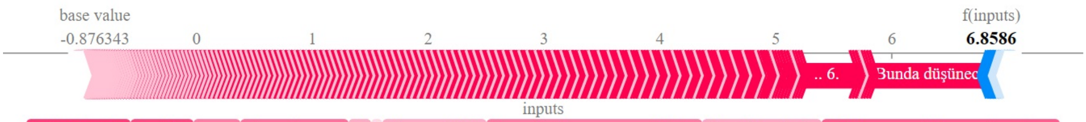
To draw a comparison between the human annotations and model explanations, it is needed to define a relation between satirical and fake content. Considering the nature of satirical news articles, it is assumed that the parts that are labelled as FAKE in the annotation are likely to contribute to the satirical meaning of the article. This can be in the form of a fake person, a fake quote, or a fake event.

Similarly, the parts that are annotated to be REAL are less likely to contribute to the

Will Smith'in dün gece düzenlenen 94. Oscar ödül töreninde sunucu Chris Rock'a attığı tokatın yankıları sürerken, gözler bu konuda henüz görüşünü bildirmemiş son sosyal medya kullanıcısı olan size çevrildi. Tokatın üzerinden 24 saatten uzun bir süre geçmesine rağmen hala Will Smith haklı mı yoksa tam bir barzo mu, Chris Rock ayıp mı etti yoksa müthiş bir beyefendilik örneği mi sergiledi soruları hala net bir yanıt bulamazken, olayın açıklığa kavuşturulup kamuoyu vicdanının rahatlayabilmesi için bir noktada artık sizin de görüşünüzü bildirmeniz gerekiyor. Şu ana dek yaklaşık 2.4 milyar kişinin tarafını seçtiği olayla ilgili olarak son derece kritik fikrinizi açıklamadan önce bilmeniz gereken önemli bilgiler ise şöyle:

1. Will Smith'in karısı kanser değilmiş. Saçkıran mı ne öyle dandik bir hastalık yüzünden saçını kazıtmış
2. Evet hakkaten vurmuş. Ama yumruk değil tokat
3. "Toksik maskülinite" kalıbını cümle içinde kullanırken dikkat edin. Yanlış yazan çok var.
4. Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı
5. Olay kurgu değil. Ama ola dabilir. Ya da yok ya değil...
6. Tuvalet kağıdında KDV oranı %8'e indirildi (Belki bir faydası olur)
7. Chris Rock haklı. Bunda düşünecek bir şey yok

Figure 5.3: Human annotation for news article (B)



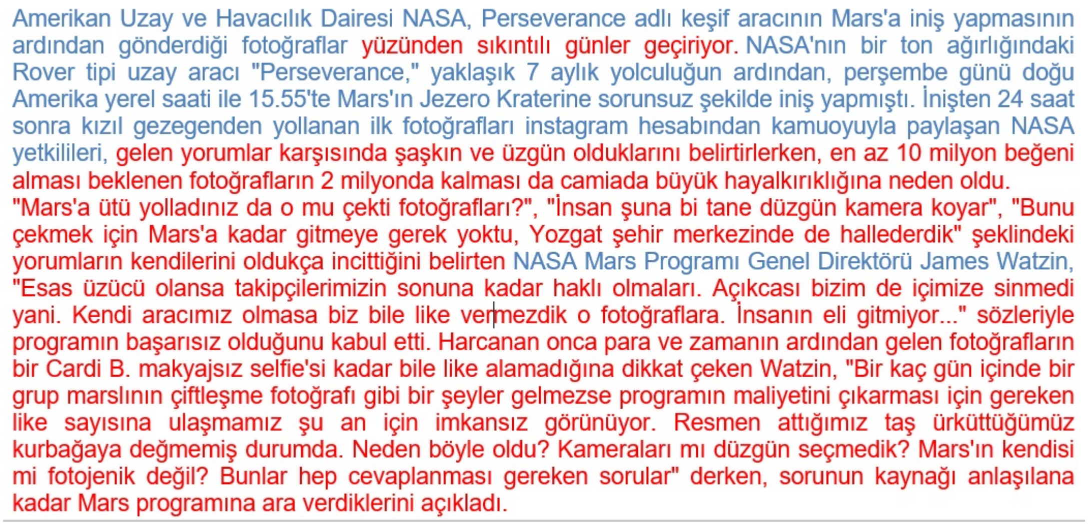
Will Smith'in dün gece düzenlenen 94. Oscar ödül töreninde sunucu Chris Rock'a attığı tokatın yankıları sürerken, gözler bu konuda henüz görüşünü bildirmemiş son sosyal medya kullanıcısı olan size çevrildi. Tokatın üzerinden 24 saatten uzun bir süre geçmesine rağmen hala Will Smith haklı mı yoksa tam bir barzo mu, Chris Rock ayıp mı etti yoksa müthiş bir beyefendilik örneği mi sergiledi soruları hala net bir yanıt bulamazken, olayın açıklığa kavuşturulup kamuoyu vicdanının rahatlayabilmesi için bir noktada artık sizin de görüşünüzü bildirmeniz gerekiyor. Şu ana dek yaklaşık 2.4 milyar kişinin tarafını seçtiği olayla ilgili olarak son derece kritik fikrinizi açıklamadan önce bilmeniz gereken önemli bilgiler ise şöyle: 1. Will Smith'in karısı kanser değilmiş. Saçkıran mı ne öyle dandik bir hastalık yüzünden saçını kazıtmış 2. Evet hakkaten vurmuş. Ama yumruk değil tokat 3. "Toksik maskülinite" kalıbını cümle içinde kullanırken dikkat edin. Yanlış yazan çok var. 4. Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı 5. Olay kurgu değil. Ama ola dabilir. Ya da yok ya değil... 6. Tuvalet kağıdında KDV oranı %8'e indirildi (Belki bir faydası olur) 7. Chris Rock haklı. Bunda düşünecek bir şey yok mk

Figure 5.4: SHAP output for news article (B)

overall satire in the text. For example, the event described in an article may be real, therefore it can be annotated as `REAL`, but there may be a fake quote in the rest of the article that contributes to the satirical meaning.

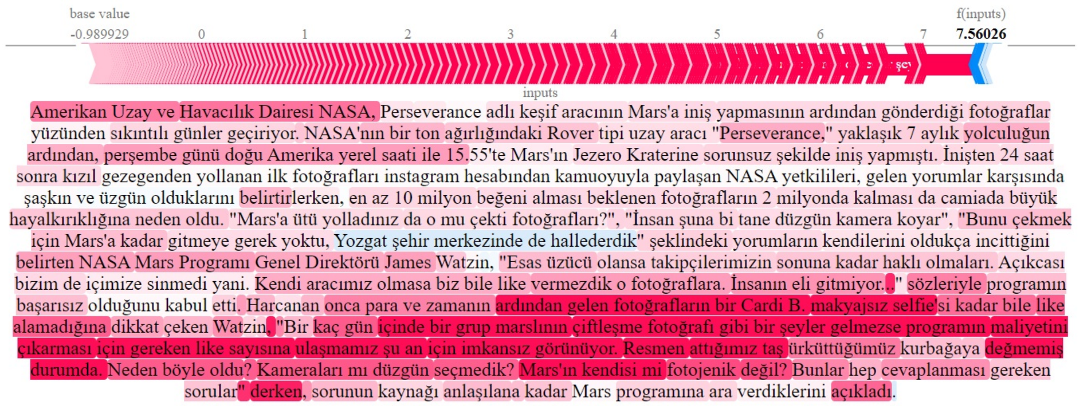
Following these parallels, three articles labelled A, B, and C, are compared in terms of their annotations and explanations. The full article texts with human annotations and English translations can be found in the Appendix A.

According to the Figures 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6, it can be seen that the SHAP output and the human annotation overlaps for most of the red highlights, meaning that the expected match between `SATIRICAL` and `FAKE` annotations is observed. On the other hand, this seems not to be the case for blue highlights, i.e. for the parts that are



Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor. NASA'nın bir ton ağırlığındaki Rover tipi uzay aracı "Perseverance," yaklaşık 7 aylık yolculuğun ardından, perşembe günü doğu Amerika yerel saati ile 15.55'te Mars'ın Jezero Kraterine sorunsuz şekilde iniş yapmıştı. İnişten 24 saat sonra kızıl gezegenden yollanan ilk fotoğrafları instagram hesabından kamuoyuyla paylaşan NASA yetkilileri, gelen yorumlar karşısında şaşkın ve üzgün olduklarını belirtirlerken, en az 10 milyon beğeni alması beklenen fotoğrafların 2 milyonda kalması da camiada büyük hayalkırıklığına neden oldu.
"Mars'a ütü yolladınız da o mu çekti fotoğrafları?", "İnsan şuna bi tane düzgün kamera koyar", "Bunu çekmek için Mars'a kadar gitmeye gerek yoktu, Yozgat şehir merkezinde de hallederdik" şeklindeki yorumların kendilerini oldukça incittiğini belirten NASA Mars Programı Genel Direktörü James Watzin, "Esas üzücü olansa takipçilerimizin sonuna kadar haklı olmaları. Açıkcası bizim de içimize sinmedi yani. Kendi aracımız olmasa biz bile like vermezdik o fotoğraflara. İnsanın eli gitmiyor..." sözleriyle programın başarısız olduğunu kabul etti. Harcanan onca para ve zamanın ardından gelen fotoğrafların bir Cardi B. makyajsız selfie'si kadar bile like alamadığına dikkat çeken Watzin, "Bir kaç gün içinde bir grup marslının çiftleşme fotoğrafı gibi bir şeyler gelmezse programın maliyetini çıkarması için gereken like sayısına ulaşmamız şu an için imkansız görünüyor. Resmen attığımız taş ürküttüğümüz kurbağaya değmemiş durumda. Neden böyle oldu? Kameraları mı düzgün seçmedik? Mars'ın kendisi mi fotojenik değil? Bunlar hep cevaplanması gereken sorular" derken, sorunun kaynağı anlaşılana kadar Mars programına ara verdiklerini açıkladı.

Figure 5.5: Human annotation for news article (C)



Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor. NASA'nın bir ton ağırlığındaki Rover tipi uzay aracı "Perseverance," yaklaşık 7 aylık yolculuğun ardından, perşembe günü doğu Amerika yerel saati ile 15.55'te Mars'ın Jezero Kraterine sorunsuz şekilde iniş yapmıştı. İnişten 24 saat sonra kızıl gezegenden yollanan ilk fotoğrafları instagram hesabından kamuoyuyla paylaşan NASA yetkilileri, gelen yorumlar karşısında şaşkın ve üzgün olduklarını belirtirlerken, en az 10 milyon beğeni alması beklenen fotoğrafların 2 milyonda kalması da camiada büyük hayalkırıklığına neden oldu. "Mars'a ütü yolladınız da o mu çekti fotoğrafları?", "İnsan şuna bi tane düzgün kamera koyar", "Bunu çekmek için Mars'a kadar gitmeye gerek yoktu, Yozgat şehir merkezinde de hallederdik" şeklindeki yorumların kendilerini oldukça incittiğini belirten NASA Mars Programı Genel Direktörü James Watzin, "Esas üzücü olansa takipçilerimizin sonuna kadar haklı olmaları. Açıkcası bizim de içimize sinmedi yani. Kendi aracımız olmasa biz bile like vermezdik o fotoğraflara. İnsanın eli gitmiyor..." sözleriyle programın başarısız olduğunu kabul etti. Harcanan onca para ve zamanın ardından gelen fotoğrafların bir Cardi B. makyajsız selfie'si kadar bile like alamadığına dikkat çeken Watzin, "Bir kaç gün içinde bir grup marslının çiftleşme fotoğrafı gibi bir şeyler gelmezse programın maliyetini çıkarması için gereken like sayısına ulaşmamız şu an için imkansız görünüyor. Resmen attığımız taş ürküttüğümüz kurbağaya değmemiş durumda. Neden böyle oldu? Kameraları mı düzgün seçmedik? Mars'ın kendisi mi fotojenik değil? Bunlar hep cevaplanması gereken sorular" derken, sorunun kaynağı anlaşılana kadar Mars programına ara verdiklerini açıkladı.

Figure 5.6: SHAP output for news article (C)

annotated as REAL by the human annotators. It is observed that the model sometimes considers these parts as an indication of the SATIRICAL label or does not use those parts in the prediction at all.

A closer look at article A (Figure 5.1 and Figure 5.2) shows that the red highlights for both the human annotation and model explanation generally match, but the blue highlights of the human annotator, i.e. the parts that are annotated as real correspond to the parts that are highlighted as slightly red or neutral by the model. Ideally, we would expect neutrality or blue highlights in the corresponding parts of the SHAP output.

On the other hand, a closer examination of article B (Figure 5.3 and Figure 5.4) shows that the model and human annotator are in disagreement for most of the annotations. Even though the more saturated reds highlighted by the model match the red highlights of the annotator, that is not the case for all of the red highlights. The SHAP output even shows blue highlights in where the corresponding human annotation is red, or vice versa.

Finally, a closer look at article C (Figure 5.5 and Figure 5.6) shows that both the human annotations and the SHAP output are less continuous than the previous two comparisons. Here, the highlights mostly line up with the human annotations, but the model generally misses the sudden truth value changes in a sentence, which is expected.

Since the three examples all belong to the set of instances where the model predicts the class correctly, and the model already predicts almost every instance correctly, ideally we would expect to see a consistent overlap between the annotations and the SHAP explanations, which is not the case. This supports the claim that the model is biased somehow, and not actually learning the satire represented in the textual data.

5.5.4 Exploring LLM Reasoning by Prompt Engineering

Another approach that is newly explored in the literature is asking LLMs to reason about themselves. Following a similar approach, we used GPT-4o[3] via the ChatGPT interface [4] to decide if a given article is satirical, and if so, explain why the model *thinks* that it is satirical. [5]

It should be noted that one side of opinions on generative LLMs such as GPT is that they are not *thinking* but only repeating nonsense [62], which is closer to the standpoint of the author of this thesis. Nevertheless, for better coverage of the existing techniques, we believe that this path should also be explored while keeping the ethical and environmental concerns [62] in mind.

[3] https://openai.com/index/hello-gpt-4o/
[4] https://chatgpt.com
[5] This experiment is first done with GPT3.5, but to better discuss the state-of-the-art models, repeated later on after the release of GPT-4o.

After some experimentation, following prompt structure is used to ask for classification and explanation in Turkish:

Sana bir haber metni vereceğim, bana haberin satirik olup olmadığını söyler misin? Eğer satirik olduğunu düşünüyorsan, nerelerinin satirik olduğunu listele ve nedenlerini açıkla.
Haber metni:

Which translates to English as:

I will give you a news article, can you tell me whether the article is satirical or not? If you think that it is satirical, list the places that have satire and explain why.
Article text:

The same articles A, B, and C are given after the prompt, with no example explanations or labels. The full explanations are in the Appendix B with their translations to English. Here, the important points are summarised.

It is seen that GPT4o, being the most advanced generalised model currently available for open use, shows generally a good understanding of satire. It mostly captures the exaggerations and mockery in the language. Surprisingly, it has the contextual information to understand some of the political and financial mockery existing in Article A.

On the other hand, we see that even though some explanations it gives may be generally true for satirical reasoning, they do not apply to specific cases. For instance, in Article B, the explanation for the sentence *"Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı."* states ridicule for the emotional situation for the reason of the satire. However, the satire here is caused by the irony of the situation, where it happened to look as if Will Smith was awarded an Oscar for his act of abuse, and cried happily for his accomplishment.

Furthermore, for Article C, the main point of satire is that almost everything stated in the article is fabricated. However, GPT4o's explanations for satire are only stating the irony and humour in such events, without questioning their legitimacy. Here, we see that human reasoning easily sees that NASA officials would not make such decisions or announcements on the basis of social media likes.

Overall, we see that GPT4o is good at identifying satirical content as a whole, but we cannot always be sure that it is identified as such for the right reasons, similar to the SHAP explanations for the BERT model.

5.6 RQ2 & RQ3: Exploring Generalisability

Since computer science literature proposes similar solutions to the problem of satire, irony, and sarcasm detection, it can be hypothesised that a model capable of detecting one may perform competently for the others.

To explore this and better understand if the data and the model trained using that data are biased, two sets of experiments are explored:

- **Satire-to-Irony pipeline:** A model fine-tuned on the *Turkish Satirical News Dataset* is tested on *IronyTR* dataset
- **Irony-to-Satire pipeline:**A model fine-tuned on the *IronyTR* dataset is tested on *Turkish Satirical News Dataset*

The results are reported in Table 5.3. It can be seen that even though using the *IronyTR* dataset results in a relatively usable model with 74.50% accuracy, the model that is fine-tuned on *Turkish Satirical News Dataset* does not show the same generalisability performance. This supports the concerns raised with RQ2 and RQ3 regarding the biases of the satirical dataset and the satire detection model.

Table 5.3: *Satire-to-Irony* and *Irony-to-Satire* pipeline performances

Model	Learning rate	*Satire-to-Irony* Accuracy	*Irony-to-Satire* Accuracy
BERTurk	0.00005	38.73%	**74.50%**
BERTurk	0.00002	25.68%	54.57%

5.7 RQ3: Debiasing Pipeline

The third research question focuses on the possibility of creating better performing and debiased models even with biased datasets. The pipeline proposed here utilises prompt engineering and synthetic data generation to remove the effect of the bias coming from the heavily stylistic language of one class. The proposed pipeline is created for to improve the usability of the curated dataset, however, we believe that this pipeline can be generalised for any biased dataset by changing the prompts according to the task and bias.

5.7.1 Reducing Stylistic Bias by Prompt Engineering

Since the explanations of GPT look promising, we hypothesised that if GPT can create a non-satirical version of the satirical articles, using them in the training set may cancel the bias caused by the language of Zaytung writers.

For generating non-satirical versions of the articles, the following prompt is used:

(Prompt 1)
"Sana satirik bir haber vereceğim, adım adım bu haberdeki satirik unsurları kaldırmanı isteyeceğim. Önce bunun için haberden çıkarılması gereken cümleleri tespit et, sonra da cümleler çıkarılmış haliyle haberi tekrar yaz.
Haber metni:"

Which translates to English as:

"I will give you a satirical news article, and I will ask you to remove the satirical elements step by step. First, identify the sentences that need to be removed from the news, and then rewrite the news with those sentences removed.
Article text:"

Unfortunately, even though GPT seems to be very competent at identifying satire, it fails to remove the satirical content from the articles most of the time. As an example, consider the generated article using Article A:

"Terra Luna Coin son günlerde büyük bir değer kaybı yaşadı. Terraform Labs CEO'su

Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı da değer kaybettiğini belirtti. Bu durum, şirket çalışanları arasında moral bozukluğuna neden oldu. Kwon, kısa vadede TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini ifade etti."

Which translates to English as:

"Terra Luna Coin has experienced a significant loss in value in recent days. Terraform Labs CEO Do Kwon stated that their investigation revealed that Luna has lost value not only against the US dollar but also against the Turkish Lira (TL). This situation has caused a morale decline among company employees. Kwon expressed that in the short term, they expect the depreciation of the TL to surpass that of Luna."

Even though the model was able to understand the satire made at the expense of the Turkish Lira for the generated explanations, it was unable to remove it from the generated article.

Also, consider the generated article from Article C:

"Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor. Mars programının başarıyla sonuçlanan inişinden sonra, NASA yetkilileri beklenen ilginin aksine, sosyal medyada düşük bir etkileşimle karşılaştı. Program direktörü James Watzin, beklenen ilgiyi alamamalarının hayal kırıklığına yol açtığını ve programın başarısız olduğunu kabul etti. Sorunun kaynağını anlamak için Mars programına ara verildi."

Which translates to English as:

"The American Space and Aeronautics Agency NASA is going through difficult days due to the photos sent after the Perseverance rover successfully landed on Mars. Contrary to expectations, NASA officials encountered low engagement on social media following the successful landing of the Mars program. Program director James Watzin admitted that the lack of expected interest has led to disappointment and that the program has been deemed a failure. The Mars program has been paused to understand the source of the problem."

Here again, the satire created by the fake social media expectation narrative is still present in the generated article.

The realisation made at this point was that even though the generated content was still satirical, the style of the writing was less exaggerated. Hence the hypothesis is changed to generating new satirical articles from the existing ones in the dataset by asking GPT to remove the exaggerations and to rewrite in a more neutral tone.

For generating satirical articles with less significant stylistic features, the following prompt is used:

(Prompt 2)
"Sana bir metin vereceğim, içindeki satirik cümleleri daha düz bir dile çevirip tekrar yaz.
Haber metni:"

Which translates to English as:

"I will give you a text, and I want you to rewrite it by translating the satirical sentences into a more straightforward language.
Article text:"

Since using LLMs for data generation is nondeterministic, the quality of the generated content differs between articles for each prompt. For example, for Article D C, Prompt 2 generates the following text:

"Taraf gazetesinin "Tutsaklara karşılık Süleyman Şah Türbesi IŞİD'e verilecek" iddiası Dışişleri Bakanlığı tarafindan kesin bir dille yalanlandı. Bakanlık, Süleyman Şah Türbesi'nin yıkılarak yerine AVM yapılması konusunda IŞİD'le görüştüklerini, ancak bu görüşmenin rehinelerle ilgisi olmadığını belirtti.
Bakanlık Sözcüsü Kamuran Aydilen, "IŞİD yıkım işinde deneyimli bir ekip. Türbeyi yıkmak için onlarla görüşüyoruz ama toprak verme pazarlığı yok. Türbenin arsası çok değerli, rehineler için böyle bir pazarlık olmaz. Orada AVM yaparak ekonomiye katkı sağlamayı planlıyoruz" dedi.
Aydilen, yapılacak AVM'nin türbenin kubbesi şeklinde bir tavanı olacağını, food court, fitness salonu ve atış poligonu gibi bölümler içereceğini, ulaşımın hızlı tren

ve İDO motorları ile sağlanacağını açıkladı. Askerlerin de AVM'de özel güvenlik ve otopark görevlisi olarak istihdam edileceğini ekledi. Ayrıca, bu projeyle kimsenin kaybetmeyeceğini ve halkın ilgisini çekeceğini belirtti."

Which translates to English as:

"The claim made by Taraf newspaper that the Süleyman Shah Tomb will be handed over to ISIS in exchange for hostages has been strongly denied by the Ministry of Foreign Affairs. The ministry clarified that they are in talks with ISIS regarding the demolition of the tomb to build a shopping mall in its place, but emphasized that these discussions are unrelated to the hostages.
Ministry spokesperson Kamuran Aydilen stated, "ISIS has an experienced demolition team. We're negotiating with them to demolish the tomb, but there is no land-for-hostages deal. The land where the tomb is located is very valuable, and we wouldn't negotiate it for hostages. We plan to contribute to the economy by building a shopping mall there."
Aydilen also mentioned that the mall will have a dome-shaped ceiling inspired by the tomb's dome and will include sections such as a food court, fitness center, and shooting range. Transportation to the mall will be provided by high-speed trains and İDO ferries. He added that the soldiers stationed there will be employed as security personnel and parking attendants at the mall. Furthermore, he emphasized that this project would not cause any losses and would attract public interest."

However, for Article A, we see that the Prompt 2 is not as able to reduce the stylistic bias as Prompt 1. Using Prompt 2, following article is generated:

"Son iki günde %99.8 değer kaybı yaşayan Terra Luna Coin, 64 dolardan 0.2 dolara kadar geriledi ve kötü haberler gelmeye devam ediyor.
Bu büyük düşüşün ardından Terraform Labs CEO'su Do Kwon, bugün yaptığı açıklamada, Luna'nın sadece ABD dolarına değil, Türk Lirası'na karşı da değer kaybettiğini tespit ettiklerini söyledi. Kwon, "Bu son iki günde olmasını beklemediğimiz her şey oldu. Çok üzgünüm," dedi.
Kwon, durumu ekibe moral vermek için düzenlediği bir toplantıda fark ettiklerini belirtti. "Ekibe, 'En azından TL karşısında değer kazandık' demek istedim ama durumu kontrol etmemiştim. Asıl hata bu oldu," dedi. Grafiği açtıklarında ise acı gerçekle

yüzleştiklerini ifade etti.
"Toplantı sırasında ekibin yarısı binayı terk etti ve hala onlara ulaşamıyoruz. Kalanlar ise ofiste satılacak bir şeyler arıyor," diye konuşan Kwon, kısa vadede TL'nin değer kaybının Luna'nınkini geride bırakabileceğini belirterek, geleceğe yönelik iyimser bir mesaj vermeyi de ihmal etmedi."

Which translates to English as:

"In the past two days, Terra Luna Coin has experienced a 99.8% loss in value, dropping from $64 to $0.2, and bad news continues to emerge.
Following this major drop, Terraform Labs CEO Do Kwon stated in a recent announcement that they have found evidence that Luna has lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we didn't expect to happen in these last two days happened. I'm very sorry."
Kwon mentioned that they realized the situation during a meeting he organized to boost the team's morale. "I wanted to tell the team, 'At least we're gaining value against the TL,' but I hadn't checked the situation. That was the real mistake," he said. He also mentioned that they faced the harsh reality when they opened the chart.
"During the meeting, half of the team left the building, and we still can't reach them. The rest are just looking for things to sell in the office," Kwon said. He also noted that in the short term, the loss of value of the TL might surpass that of Luna, and he did not forget to provide an optimistic message for the future."

Since the data generation process for this study includes a human in the loop, both prompts are used to generate a total of 200 articles. Overall, the generated articles are shorter and include fewer quotes, mockery, and exaggerations than their original counterparts.

5.7.1.1 Statistics on Prompt Generated Articles

200 articles that are generated using the aforementioned 2 prompts are checked manually to see if they still have the satirical and ironic subtext present. Additionally, they are analysed in terms of the context, if the original context of the article is still understandable or some of the context has been lost during the translation process. It

is seen that:

- Out of the 200 articles, 29 of them can be labelled as `LEGITIMATE` by an unsuspecting reader with not enough knowledge on the Turkish political landscape.
- Out of that 29 articles, all 29 of them have a major loss of context compared to the original article.
- 28 of the 29 articles that are labelled as `LEGITIMATE` are generated with Prompt 1, and a loss of context and satirical subtext is understandable.

Using both prompts yields us a 14.5% error rate, which is a rate we believe is acceptable in the scope of this work. This error rate can be improved if Prompt 1 is removed from the pipeline and only Prompt 2 is utilised in the automatic generation process.

5.7.2 Pipeline Design

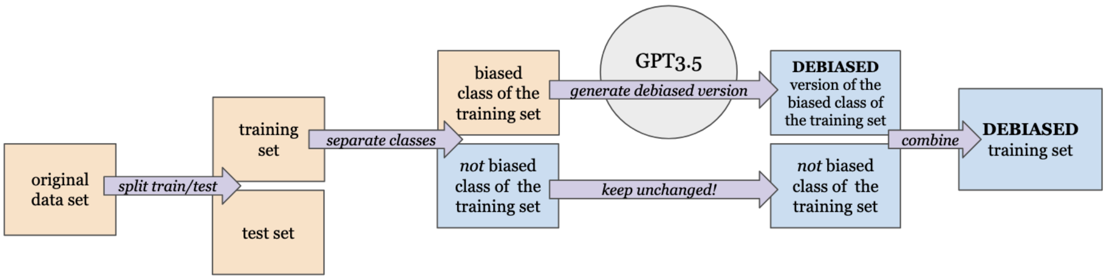


Figure 5.7: The proposed debiasing pipeline

The proposed pipeline, shown in Figure 5.7 works as follows:

- Data is separated into train and test sets
- Stylistically biased data in the train set is removed
- Removed data is used to generate more neutral and less exaggerated versions of them using GPT (or other generative LLM)
- Generated data is added back to the train set belonging to their original labels

- The classifier is trained/fine-tuned on this dataset

In this study, the data generation is done via the web interface of GPT: ChatGPT. However, the generation pipeline can easily be automatised by using the OpenAI API for GPT3.5 to fully automate the debiasing process.

5.7.3 Experiments and Results

To analyse the performance of the proposed debiasing pipeline, two sets of experiments are conducted. *BIASED* and *DEBIASED* namings are given according to the data used to fine tune the model during the training process.

BIASED pipeline is trained on 200 sampled articles from the `LEGITIMATE` class of the original dataset, and the original versions of the 200 generated articles for the `SATIRICAL` label. On the other hand, *DEBIASED* pipeline is again trained on 200 sampled articles from the `LEGITIMATE` class of the original dataset, but the 200 generated articles for the `SATIRICAL` label.

Both *BIASED* and *DEBIASED* pipelines are explored with different models and parameters. Training performance of BERTurk and multilingual BERT model are explored in the following parameter space:

- Batch size: 8
- Epoch: 2
- Learning rate: 0.00005, 0.00002

In the first set of experiments, trained model performance is tested using three different test sets with only positive labeled instances (satirical or ironic instances):

- **Zaytung:** 2002 Turkish articles labeled as `SATIRICAL` and not used in the generation/training process from the *Turkish Satirical News Dataset*.

- **the Onion:** 9000 English satirical news article headlines from the American satirical news website the Onion[6], taken from the openly available dataset OnionOrNot[7].

- **IronyTR:** 300 Turkish ironic short form texts from IronyTR dataset[8].

Tables 5.4, 5.5, and 5.6 report the performances of each experiment setting for 3 test sets as described above.

Table 5.4: *BIASED* and *DEBIASED* pipeline performances for Zaytung test set

Model	Learning rate	*BIASED* Accuracy	*DEBIASED* Accuracy
BERTurk	0.00005	**95.46%**	58.31%
BERTurk	0.00002	92.06%	47.18%
BERT multilingual	0.00005	90.96%	**83.13%**
BERT multilingual	0.00002	87.02%	36.50%

Looking at Table 5.4, a dramatic decrease in the accuracy of the model for classifying the *Turkish Satirical News Dataset* is observed when the generated data is used for training. These results seem the prove the hypothesis that the model is learning the stylistic language of the news source instead of the satirical meaning of the articles. The results show that our proposed model breaks this bias, however, for most of the models, this makes the classifier perform almost unusable. On the other hand, the multilingual BERT model with a lower learning rate is still accurate enough even after the debiasing process.

Table 5.5: *BIASED* and *DEBIASED* pipeline performances for the Onion test set

Model	Learning rate	*BIASED* Accuracy	*DEBIASED* Accuracy
BERTurk	0.00005	**38.73%**	03.74%
BERTurk	0.00002	25.68%	**70.24%**
BERT multilingual	0.00005	01.09%	00.41%
BERT multilingual	0.00002	38.24%	30.50%

[6] https://theonion.com/
[7] https://www.kaggle.com/datasets/chrisfilo/onion-or-not
[8] https://github.com/teghub/IronyTR

On Table 5.5, it can be seen that both the biased and debiased training pipelines produce unusable classification models with respect to the Onion data, with the exception of BERTurk model with 0.00002 learning rate. However, it is hard to tell if this result is reproducable since BERTurk model have no prior knowledge of the English language. It should be noted that there is a general trend of decrease for the other models, and this may be because of the possible common words that exist in both the Zaytung and the Onion articles (common words or names in English and Turkish), which is worth exploring.

Table 5.6: *BIASED* and *DEBIASED* pipeline performances for IronyTR test set

Model	Learning rate	*BIASED* Accuracy	*DEBIASED* Accuracy
BERTurk	0.00005	**77.10%**	**100.00%**
BERTurk	0.00002	72.05%	21.21%
BERT multilingual	0.00005	**77.10%**	99.66%
BERT multilingual	0.00002	52.53%	69.02%

Table 5.6 shows the accuracy scores for the models when tested with the ironic set of the IronyTR dataset. Since the debiasing process claims to generalise the model with respect to the detection of satire and irony, an increase in the performance of classification after the debiasing process is expected. As such, the expected results are observed. With the exception of BERTurk with 0.00002 learning rate, we see a consistent increase in the models for the accuracy, even achieving 100%.

Results reported up to here focuses on the effects of the debiasing process on satirical and ironic data. To further understand the effects of the debiasing process, another set of experiments are conducted with testing sets having both non-satirical/non-ironic and satirica/ironic instances:

- **Balanced Zaytung + Legitimate:** 2002 Turkish articles labeled as `SATIRICAL` and 2002 Turkish articles labeled as `LEGITIMATE`, not used in the generation/-training process from the *Turkish Satirical News Dataset*.
- **Balanced the Onion + HuffPost:** A fairly balanced set of 29000 English news article headlines from the American satirical news website the Onion[9]

[9] https://theonion.com/

and HuffPost[10], taken from the openly available News Headlines Dataset For Sarcasm Detection[11].

- **Balanced IronyTR:** 600 Turkish short form texts from IronyTR dataset[12], both ironic and non-ironic.

Tables 5.7, 5.8, and 5.9 present the performances of the models on the balanced test sets listed above. Since the balanced sets have both positive and negative labels, performances are presented with precision, recall, accuracy, and F1 scores. Results are mainly compared with respect to their F1 scores.

Table 5.7: *BIASED* and *DEBIASED* pipeline performances for Balanced Zaytung test set

Model	Learning rate	Accuracy	Precision	Recall	F1
BIASED					
BERTurk	0.00005	98.69%	98.70%	98.69%	**98.69%**
BERTurk	0.00002	95.62%	95.85%	95.62%	95.62%
BERT multilingual	0.00005	97.41%	97.53%	97.41%	97.41%
BERT multilingual	0.00002	94.72%	95.14%	94.72%	94.71%
DEBIASED					
BERTurk	0.00005	80.81%	85.19%	80.81%	**80.25%**
BERTurk	0.00002	71.85%	79.27%	71.85%	70.08%
BERT multilingual	0.00005	44.93%	34.15%	44.93%	33.53%
BERT multilingual	0.00002	64.20%	68.57%	64.20%	62.17%

Table 5.7 presents the performance scores for the Balanced Zaytung test set. Similar to the Zaytung test set presented in Table 5.4, a consistent decrease is observed. However, when the set is balanced, the results show us that the performance of the model is still competent and usable, with the exception of the multilingual BERT model with learning rate 0.00005.

It can be seen from Table 5.8 that even though the models are more usable on the

[10] https://www.huffpost.com/

[11] https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection/data

[12] https://github.com/teghub/IronyTR

Table 5.8: *BIASED* and *DEBIASED* pipeline performances for Balanced the Onion test set

Model	Learning rate	Accuracy	Precision	Recall	F1
BIASED					
BERTurk	0.00005	51.98%	49.55%	51.98%	38.46%
BERTurk	0.00002	48.80%	48.29%	48.80%	47.81%
BERT multilingual	0.00005	51.45%	40.41%	51.45%	34.99%
BERT multilingual	0.00002	52.96%	52.57%	52.96%	**51.66%**
DEBIASED					
BERTurk	0.00005	49.78%	49.30%	49.78%	48.75%
BERTurk	0.00002	52.49%	52.76%	52.49%	**52.49%**
BERT multilingual	0.00005	52.36%	27.42%	52.36%	34.37%
BERT multilingual	0.00002	45.82%	46.20%	45.82%	44.28%

Balanced the Onion set than the models presented in Table 5.5, they are still not competent, both for the *BIASED* and *DEBIASED* training pipelines. However, one thing to look into is even though the debiasing process results in 5-10% increase in the performance of the BERTurk results, the opposite is observed for the multilingual BERT model. Again, this may be because of the possible common words that exist in both the Zaytung and the Onion articles (common words or names in English and Turkish).

As the last results in this set of experiments, results for the Balanced IronyTR test set are reported in Table 5.9. Similar to the trend in Table 5.6, classification performance improves after the debiasing process with Zaytung data, as expected.

Finally, *COMBINED* pipelines combining both the biased and debiased data for the finetuning process are explored. This set of experiments aims to see if combining original and debiased data yields a middle ground between generalisability and classification performance. The training set used in the finetuning process includes 100 of the 200 generated Zaytung articles, and 100 random samples taken from the remaining original Zaytung articles. The test sets are as follows:

- **Balanced Zaytung + Legitimate:** 1902 Turkish articles labeled as `SATIRICAL`

Table 5.9: *BIASED* and *DEBIASED* pipeline performances for Balanced IronyTR test set

Model	Learning rate	Accuracy	Precision	Recall	F1
BIASED					
BERTurk	0.00005	51.93%	52.76%	51.93%	47.12%
BERTurk	0.00002	47.07%	46.23%	47.07%	43.55%
BERT multilingual	0.00005	53.43%	55.83%	53.43%	48.60%
BERT multilingual	0.00002	58.12%	61.30%	58.12%	**55.18%**
DEBIASED					
BERTurk	0.00005	56.11%	59.01%	56.11%	52.58%
BERTurk	0.00002	54.27%	57.22%	54.27%	48.62%
BERT multilingual	0.00005	55.11%	58.63%	55.11%	49.62%
BERT multilingual	0.00002	60.30%	60.37%	60.30%	**60.20%**

and 1902 Turkish articles labeled as `LEGITIMATE`, not used in the generation/-training process from the *Turkish Satirical News Dataset*.

- **Balanced the Onion + HuffPost:** A fairly balanced set of 29000 English news article headlines from the American satirical news website the Onion[13] and HuffPost[14], taken from the openly available News Headlines Dataset For Sarcasm Detection[15].

- **Balanced IronyTR:** 600 Turkish short form texts from IronyTR dataset[16], both ironic and non-ironic.

Results for the *COMBINED* pipelines are presented in Table 5.10. Looking at the Balanced Zaytung + Legitimate tests and comparing them with the *BIASED* and *DEBIASED* results reported in Table 5.7, it is seen that the combining equal number of instances of biased and debiased samples for training resulted in a performance that is closer to the *BIASED* pipeline.

[13] https://theonion.com/
[14] https://www.huffpost.com/
[15] https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection/data
[16] https://github.com/teghub/IronyTR

To make sure whether this is a failure for debiasing or not, *COMBINED* pipeline performances for the Onion+HuffPost and Balanced IronyTR should be compared with the *BIASED* and *DEBIASED* pipeline performances reported in Tables 5.8 and 5.9 respectively. Upon comparison, it is seen that for both test sets, the generalisation performance of the model has dropped almost to the *BIASED* pipelines' levels. This unfortunately shows that equally combining biased and debiased instances almost takes us back to the original performance.

Table 5.10: *COMBINED* pipeline performances for balanced test sets

Model	Learning rate	Accuracy	Precision	Recall	F1
Zaytung+Legitimate					
BERTurk	0.00005	97.35%	97.39%	97.35%	**97.35%**
BERTurk	0.00002	97.51%	97.53%	97.51%	97.51%
BERT multilingual	0.00005	92.77%	92.77%	92.77%	92.77%
BERT multilingual	0.00002	92.66%	92.93%	92.66%	92.65%
the Onion+HuffPost					
BERTurk	0.00005	50.96%	53.13%	50.96%	**48.14%**
BERTurk	0.00002	50.05%	47.77%	50.05%	43.50%
BERT multilingual	0.00005	48.78%	49.58%	48.78%	47.83%
BERT multilingual	0.00002	52.17%	50.86%	52.17%	41.68%
Balanced IronyTR					
BERTurk	0.00005	51.09%	70.13%	59.46%	36.39%
BERTurk	0.00002	54.77%	55.43%	58.29%	53.09%
BERT multilingual	0.00005	59.46%	63.54%	51.09%	**56.37%**
BERT multilingual	0.00002	58.29%	61.59%	54.77%	55.32%

Overall, it is observed that the debiasing process using the Zaytung data can successfully reduce the bias of the model against the stylistic language of the Zaytung writers, and make the model more generalisable for being used in other datasets after training. Different combinations of biased and debiased data may be further explored, but initial experiments with equal combinations show that it almost reverts the effects of the debiasing process.

5.8 Discussion

The problem of satire detection demands a human in the loop by its nature since the labelling process cannot be automated. The only automatisation possible is finding a satirical resource such as Zaytung and assuming all scraped content is satirical by default. Unfortunately, this causes the data to be biased stylistically and trickles down this bias to the model where the model learns to identify the style of the writing instead of the satire, as shown in this chapter.

This chapter of the thesis proposes a debiasing pipeline utilising LLM-based text generation in ethical limits. We show that generating data that is stylistically neutral to replace the biased data in the training set decreases the model performance significantly. However, additional experimentation is needed to see if this method improves the generalisability of the model, and if it is generalisable itself as a debiasing method for different language tasks. Yet, we believe that the results look promising.

When using LLMs, specifically generative LLMs, ethical and environmental concerns should always be kept in mind. Generating textual data is also convoluted ethically, and should not be taken lightly. We believe that LLM-generated data should not be contextualised as if a real human has generated that content.

CHAPTER 6

CONCLUSIONS

This thesis explores the irony and satire detection problems, which are important problems for fighting misinformation online, in different contexts. Separated into three chapters, three main problems are covered: irony detection, satire detection, and debiasing of models that are trained with stylistically biased text.

In Chapter 3, the problem of irony detection is studied in detail. Firstly, *IronyTR: Turkish Irony Dataset* is curated from online microblog posts. The balanced dataset consists of 600 instances that are labelled by a majority voting of 7 annotators.

The rest of the chapter thoroughly explores the effects of different sets of features on the performance of irony detection models. Using BERT, SVM, DT, and NB based classifiers and the *IronyTR* dataset, different subsets of lexical, syntactic, polarity-based and graph-based features are compared. It is reported that polarity score and graph-based features improve the performance of traditional classifiers. On the other hand, BERT outperforms all pipelines and gives more promising results with no feature engineering, but only using the tokenised data instances.

In Chapter 4, in a parallel manner with Chapter 3, the problem of satire detection is studied, with a specific focus on news articles. Firstly, the initial version of the *Turkish Satirical News Dataset* is curated from online news sites, with satirical news articles taken from Zaytung[1].

The rest of the chapter, again in a similar manner to Chapter 3, explores the effects of different sets of features on the performance of the satirical news detection models. To understand the impact of the bodies and titles of the articles, two sets of experiments

[1] https://zaytung.com/

are conducted, with two different representations for the articles. It is reported that when the articles are represented with their titles and the body text, and all extracted features are used, the SVM-based model shows competitive performance. However, once again we see that the BERT-based model outperforms all models with no feature engineering.

On the other hand, the almost-perfect performance of the BERT model, as well as the very high performance of other models in several other experiment settings raise questions about the quality of the dataset. A possibility of bias in the dataset and consequently the trained models is stated as a conclusion for this chapter.

Chapter 5 builds upon the two previous chapters and works on exploring the bias claims by utilising the works in both explainability and bias/debiasing literature.

Firstly, improvements are made to the *Turkish Satirical News Dataset* by removing articles published before 2014 and using a more reputable source for legitimate news. Then, the work is structured around three research questions:

- **RQ1:** Can we analyse our datasets to see if they are biased stylistically?
- **RQ2:** Can we explain the decisions of our models to discuss if they are biased?
- **RQ3:** Can we design a specialised pipeline to train an unbiased model with bias-prone data?

The answer to RQ1 is straightforward. Even a basic statistical analysis of *Turkish Satirical News Dataset* shows that there are complexity differences between classes. Using TF-IDF to extract the top 10 words for each class, it is also seen that there are apparent differences in the stylistic language of each class. A better way to further investigate this can be using the NELA features library [50, 49], which currently does not directly support Turkish but can be adapted with the help of linguistic experts. Another approach may be to use more surface, lexical, or syntactic level features covered in the literature[63], which are more adaptable to Turkish language.

The most robust way to debiasing a dataset is by collecting unbiased data in the first place. Unfortunately, collecting unbiased data requires either a wide set of satirical publications, or human annotators and data collectors to actively look for satirical content. In the current landscape of satirical content in Turkish, the only source for satirical content that we can safely assume will always publish satirical content is the online newspaper Zaytung. Hence, we focus on debiasing the models instead of the data.

The answer to RQ2 is more complicated. One way to understand if a model overfits is to test the model with related but unseen data, which we do not have such a dataset. Another approach is using explainability methods to see what features affect the decision of the model.

We utilise two different explainability approaches: using SHAP [46] to generate explanations for the decisions of the BERT model, and using GPT to reason about its classification of a given instance without any fine-tuning. Then, the explanations are compared with the annotations of a human annotator. It is seen that the BERT model has a vague understanding of satire but its classifications are not directly connected to the satirical parts of the data instances. On the other hand, we see that GPT has a better understanding of satire in general, but fails to capture the reason for satire when more human-like context information is needed. Overall, it is seen that the BERT model that uses the dataset during the training phase is biased.

Finally, RQ3 is explored by the design of a debiasing pipeline. The proposed pipeline works by replacing the training data from the class with the stylistic bias with a less biased version of each instance generated by a generative LLM. Several approaches and prompts for this generation process are explored and the best-performing approach is chosen with human judgement.

The debiasing performance of the pipeline is reported with 200 generated instances, and different test sets. Overall, it is observed that the debiasing process using the Zaytung data can successfully reduce the bias of the model against the stylistic language of the Zaytung writers, and make the model more generalisable for being used in other datasets after training. We believe that this debiasing approach is a step in the right direction, but still very much open to improvement.

Last but not least, it should be mentioned that using generative LLMs such as GPT variants is a subject that comes with an ethical burden. These ethical concerns include the environmental impact of these models and the violation of intellectual property. We believe that LLM-generated text should never replace real textual content generated by real humans, and when using these models, environmental impact should always be considered.

REFERENCES

[1] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, “Misinformation in Social Media: Definition, Manipulation, and Detection,” *ACM SIGKDD Explorations Newsletter*, vol. 21, pp. 80–90, Nov. 2019.

[2] E. Aïmeur, S. Amri, and G. Brassard, “Fake news, disinformation and misinformation in social media: a review,” *Social Network Analysis and Mining*, vol. 13, p. 30, Feb. 2023.

[3] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, M. Sandhu, R. Sang, R. Scholz-Bright, A. T. Welch, A. G. Wolff, A. Zhou, and B. Nyhan, “Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media,” *Political Behavior*, vol. 42, pp. 1073–1095, Dec. 2020.

[4] J. Ling and R. Klinger, “An Empirical, Quantitative Analysis of the Differences Between Sarcasm and Irony,” in *The Semantic Web* (H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, and C. Lange, eds.), vol. 9989, pp. 203–216, Cham: Springer International Publishing, 2016. Series Title: Lecture Notes in Computer Science.

[5] F. Barbieri, H. Saggion, and a. F. Ronzano, “Modelling sarcasm in twitter, a novel approach,” *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 50–58, 2014.

[6] P. Carvalho, B. Martins, H. Rosa, S. Amir, J. Baptista, and M. J. Silva, “Situational irony in farcical news headlines,” in *International Conference on Computational Processing of the Portuguese Language*, pp. 65–75, Springer, 2020.

[7] C. Van Hee, E. Lefever, and V. Hoste, “Semeval-2018 task 3: Irony detection in english tweets,” in *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 39–50, 2018.

[8] C. Van Hee, E. Lefever, and V. Hoste, “Exploring the realization of irony in Twitter data,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 1794–1799, 2016.

[9] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y-Gómez, and J. E. Medina Pagola, “Overview of the task on irony detection in Spanish variants,” in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), Co-Located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. Org*, vol. 2421, pp. 229–256, 2019.

[10] R. Xiang, X. Gao, Y. Long, A. Li, E. Chersoni, Q. Lu, and C.-R. Huang, “Ciron: A new benchmark dataset for Chinese irony detection,” in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5714–5720, European Language Resources Association, 2020.

[11] A. U. Ozturk, Y. Cemek, and P. Karagoz, “IronyTR: Irony detection in turkish informal texts,” *International Journal of Intelligent Information Technologies (IJIIT)*, vol. 17, no. 4, pp. 1–18, 2021.

[12] B. Ghanem, J. Karoui, F. Benamara, P. Rosso, and V. Moriceau, “Irony detection in a multilingual context,” in *European Conference on Information Retrieval*, pp. 141–149, Springer, 2020.

[13] A. Joshi, P. Bhattacharyya, and M. J. Carman, “Automatic sarcasm detection: A survey,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–22, 2017.

[14] U. B. Baloglu, B. Alatas, and H. Bingol, “Assessment of supervised learning algorithms for irony detection in online social media,” *1st International Informatics and Software Engineering Conference (UBMYK)*, pp. 1–5, 2019.

[15] K. Buschmeier, P. Cimiano, and R. Klinger, “An impact analysis of features in a classification approach to irony detection in product reviews,” in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 42–49, 2014.

[16] E. W. Pamungkas and V. Patti, “# NonDicevoSulSerio at SemEval-2018 task 3: Exploiting emojis and affective content for irony detection in english tweets,”

in *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 649–654, 2018.

[17] C. Van Hee, E. Lefever, and V. Hoste, "Monday mornings are my fave:)# not exploring the automatic recognition of irony in english tweets," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2730–2739, 2016.

[18] U. Ahmed, L. Zafar, F. Qayyum, and M. A. Islam, "Irony detection in english tweets using word graph," *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 581–586, 2018.

[19] Y. Y. Tan, C.-O. Chow, J. Kanesan, J. H. Chuah, and Y. Lim, "Sentiment analysis and sarcasm detection using deep multi-task learning," *Wireless personal communications*, vol. 129, no. 3, pp. 2213–2237, 2023.

[20] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang, "THU-NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning," *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 51–56, 2018.

[21] S. Zhang, X. Zhang, J. Chan, and P. Rosso, "Irony detection via sentiment-based transfer learning," *Information Processing & Management*, vol. 56, no. 5, pp. 1633–1644, 2019.

[22] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella, "Transformer-based models for multimodal irony detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 7399–7410, 2023.

[23] Y. Lin, Y. Xia, and Y. Long, "Augmenting emotion features in irony detection with large language modeling," *arXiv preprint arXiv:2404.12291*, 2024.

[24] S. Frenda and V. Patti, "Computational models for irony detection in three spanish variants.," in *IberLEF@ SEPLN*, pp. 297–309, 2019.

[25] O. Dulger, "Turkce metinlerde ironi tespiti (irony classification in turkish text)," *Ulusal Yazılım Mühendisliği Sempozyumu (UYMS)*, 2018.

[26] H. Taslioglu and P. Karagoz, "Irony detection on microposts with limited set of features," in *Proceedings of the Symposium on Applied Computing*, pp. 1076–1081, 2017.

[27] Y. Cemek, C. Cidecio, A. U. Ozturk, R. F. Cekinel, and P. Karagoz, "Turkce resmi olmayan metinlerde ironi tespiti icin sinirsel yontemlerin incelenmesi (Investigating the neural models for irony detection on turkish informal texts)," *IEEE Sinyal Isleme ve Iletisim Uygulamalari Kurultayi (SIU2020)*, 2020.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[29] W. Y. Wang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, 2017.

[30] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, "Fa-kes: A fake news dataset around the syrian war," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 573–582, 2019.

[31] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *International Workshop on Combating On-Line Hostile Posts in Regional Languages during Emergency Situation*, pp. 21–29, Springer, 2021.

[32] F. Yang, A. Mukherjee, and E. Dragut, "Satirical news detection and analysis using attention mechanism and linguistic features," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1979–1989, 2017.

[33] S. De Sarkar, F. Yang, and A. Mukherjee, "Attending sentences to detect satirical fake news," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3371–3380, 2018.

[34] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and VERification," in *NAACL-HLT*, 2018.

[35] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[36] R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, pp. 13–18, 2023.

[37] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, *et al.*, "Fake news vs satire: A dataset and analysis," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 17–21, 2018.

[38] J. F. Low, B. C. Fung, F. Iqbal, and S.-C. Huang, "Distinguishing between fake news and satire with transformers," *Expert Systems with Applications*, vol. 187, p. 115824, 2022.

[39] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pp. 7–17, 2016.

[40] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[41] G. K. Koru and Ç. Uluyol, "Detection of turkish fake news from tweets with bert models," *IEEE Access*, 2024.

[42] S. G. Taskin, E. U. Kucuksille, and K. Topal, "Detection of turkish fake news in twitter with machine learning algorithms," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2359–2379, 2022.

[43] U. Mertoğlu and B. Genç, "Automated fake news detection in the age of digital libraries," *Information Technology and Libraries*, vol. 39, no. 4, 2020.

[44] A. Onan and M. A. Toçoğlu, "Satire identification in turkish news articles based on ensemble of classifiers," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 2, pp. 1086–1106, 2020.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[46] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[47] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 395–405, 2019.

[48] E. B. Buyukbas, A. H. Dogan, A. U. Ozturk, and P. Karagoz, "Explainability in irony detection," in *Big Data Analytics and Knowledge Discovery: 23rd International Conference, DaWaK 2021, Virtual Event, September 27–30, 2021, Proceedings 23*, pp. 152–157, Springer, 2021.

[49] B. D. Horne, W. Dron, S. Khedr, and S. Adali, "Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news," in *Companion Proceedings of the The Web Conference 2018*, pp. 235–238, 2018.

[50] B. D. Horne, J. Nørregaard, and S. Adali, "Robust fake news detection over time and attack," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 1, pp. 1–23, 2019.

[51] C. Qian, F. Feng, L. Wen, C. Ma, and P. Xie, "Counterfactual inference for text classification debiasing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5434–5445, 2021.

[52] I. B. Schlicht, D. Altiok, M. Taouk, and L. Flek, "Pitfalls of conversational llms on news debiasing," *arXiv preprint arXiv:2404.06488*, 2024.

[53] H. Xu, E. Santus, A. Laszlo, and C.-R. Huang, "LLT-PolyU: Identifying sentiment intensity in ironic tweets," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 673–678, 2015.

[54] A. G. Vural, B. B. Cambazoglu, P. Senkul, and Z. O. Tokgoz, "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish," in *Computer and Information Sciences III*, pp. 437–445, Springer, 2013.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[56] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.

[57] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *Journal of medical systems*, vol. 26, no. 5, pp. 445–463, 2002.

[58] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial Naïve Bayes classifier to text classification," in *Advanced Multimedia and Ubiquitous Engineering*, pp. 347–352, Springer, 2017.

[59] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[60] S. Schweter, "BERTurk - BERT models for turkish." Zenodo, Apr. 2020.

[61] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[62] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

[63] G. Verma and B. V. Srinivasan, "A lexical, syntactic, and semantic perspective for understanding style in text," *arXiv preprint arXiv:1909.08349*, 2019.

APPENDICES

A Human Annotated Satirical News

A.1 Article A - Annotation

LUNA'da Kriz Büyüyor: TL'ye Karşı Bile Değer Kaybettiği Ortaya Çıktı...
2022-05-13

Yaşadığı %99.8'lik değer kaybıyla 2 gün içerisinde 64 dolar seviyesinden 0.2 dolar seviyesine gerileyen Terra Luna Coin'den kötü haberler gelmeye devam ediyor.

Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından bugün bir açıklama yayınlayan Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken, "Şu son 2 günde olmaz dediğimiz ne varsa hepsi oldu. Çok üzgünüm" ifadelerine yer verdi.

Durumun ekibe moral vermek için yaptığı bir toplantıda ortaya çıktığını belirten Kwon "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım. Öncesinde 'nasılsa TL'den de daha kötü durumda değilizdir' diye bakma gereği duymamıştım. Esas hata o oldu" derken, grafiği ekranda açmasıyla birlikte acı gerçeği fark ettiklerini dile getirdi.

"O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz. Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten." sözleriyle Terra Labs'daki son durumu da aktaran deneyimli CEO, kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek ileriye dönük iyimser mesajlar vermeyi de ihmal etmedi.

Figure 6.1: Human annotated full article A

A.2 Article A - Translation

Crisis Deepens for LUNA: It Has Even Lost Value Against the Turkish Lira…

Following a 99.8% loss in value, where Terra Luna Coin dropped from $64 to $0.20 within two days, bad news continues to emerge.

After this dramatic drop that shook the cryptocurrency market, Do Kwon, CEO of Terraform Labs, released a statement today. He noted that their investigation revealed Luna had lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we thought couldn't happen in the last two days has happened. I am very sorry."

Kwon explained that the realization came during a meeting he held to boost the team's morale, saying, "I gathered the team to give a motivational speech along the lines of 'Look, it's not so bad. At least we are still gaining value in TL terms.' I hadn't felt the need to check if we were worse off compared to TL. That was the real mistake." He shared that the painful truth became apparent when they opened the graph on the screen.

"At that moment, half the team left the building. We still can't reach them. The remaining team members are just looking around the office to see if there's anything left to sell," Kwon described the current situation at Terra Labs. The experienced CEO also mentioned that they expect the depreciation of the TL to eventually surpass Luna, whether due to some absurd move by Turkey or the Central Bank of the Republic of Turkey depleting its remaining foreign currency reserves. He added that he has not neglected to give optimistic messages for the future.

A.3 Article B - Annotation

Uyarı: Will Smith'in Attığı Tokatla İlgili Görüş Bildirmeyen Bir Tek Siz Kaldınız...
2022-03-29

Will Smith'in dün gece düzenlenen 94. Oscar ödül töreninde sunucu Chris Rock'a attığı tokatın yankıları sürerken, gözler bu konuda henüz görüşünü bildirmemiş son sosyal medya kullanıcısı olan size çevrildi. Tokatın üzerinden 24 saatten uzun bir süre geçmesine rağmen hala Will Smith haklı mı yoksa tam bir barzo mu, Chris Rock ayıp mı etti yoksa müthiş bir beyefendilik örneği mi sergiledi soruları hala net bir yanıt bulamazken, olayın açıklığa kavuşturulup kamuoyu vicdanının rahatlayabilmesi için bir noktada artık sizin de görüşünüzü bildirmeniz gerekiyor.

Şu ana dek yaklaşık 2.4 milyar kişinin tarafını seçtiği olayla ilgili olarak son derece kritik fikrinizi açıklamadan önce bilmeniz gereken önemli bilgiler ise şöyle:

1. Will Smith'in karısı kanser değilmiş. Saçkıran mı ne öyle dandik bir hastalık yüzünden saçını kazıtmış
2. Evet hakkaten vurmuş. Ama yumruk değil tokat
3. "Toksik maskülinite" kalıbını cümle içinde kullanırken dikkat edin. Yanlış yazan çok var.
4. Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı
5. Olay kurgu değil. Ama ola dabilir. Ya da yok ya değil...
6. Tuvalet kağıdında KDV oranı %8'e indirildi (Belki bir faydası olur)
7. Chris Rock haklı. Bunda düşünecek bir şey yok

Figure 6.2: Human annotated full article B

A.4 Article B - Translation

Warning: You're the Only One Who Hasn't Weighed In on Will Smith's Slap...

As the repercussions of Will Smith's slap on comedian Chris Rock at the 94th Oscars last night continue to ripple, all eyes have turned to you, the last social media user who has yet to express an opinion on the matter. Over 24 hours have passed since the slap, and the questions of whether Will Smith was justified or simply out of line, and whether Chris Rock was rude or demonstrated exemplary gentlemanliness, still lack clear answers. For the sake of clarifying the situation and easing the public conscience, it's time for you to share your view.

Before you disclose your crucial opinion on this event, which has involved approximately 2.4 billion people choosing sides, here are some important details you need to know:

1. Will Smith's wife is not suffering from cancer. She shaved her head due to some trivial disease like alopecia. 2. Yes, he really did hit him. But it was a slap, not a punch. 3. Be careful when using the term "toxic masculinity" in a sentence. Many people spell it wrong. 4. After delivering the slap, Will Smith went on to cry and then won the Oscar. 5. The incident is not staged. But it could be. Or not... who knows. 6. The VAT rate on toilet paper has been reduced to 8% (Maybe this will help). 7. Chris Rock is right. There's nothing more to think about.

A.5 Article C - Annotation

NASA, Fotoğrafların Beklenen Like Sayısına Ulaşamaması Nedeniyle Mars Programını Sonlandırdığını Açıkladı...
2021-02-20

Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor.

NASA'nın bir ton ağırlığındaki Rover tipi uzay aracı "Perseverance," yaklaşık 7 aylık yolculuğun ardından, perşembe günü doğu Amerika yerel saati ile 15.55'te Mars'ın Jezero Kraterine sorunsuz şekilde iniş yapmıştı. İnişten 24 saat sonra kızıl gezegenden yollanan ilk fotoğrafları instagram hesabından kamuoyuyla paylaşan NASA yetkilileri, gelen yorumlar karşısında şaşkın ve üzgün olduklarını belirtirlerken, en az 10 milyon beğeni alması beklenen fotoğrafların 2 milyonda kalması da camiada büyük hayalkırıklığına neden oldu.

"Mars'a ütü yolladınız da o mu çekti fotoğrafları?", "İnsan şuna bi tane düzgün kamera koyar", "Bunu çekmek için Mars'a kadar gitmeye gerek yoktu, Yozgat şehir merkezinde de hallederdik" şeklindeki yorumların kendilerini oldukça incittiğini belirten NASA Mars Programı Genel Direktörü James Watzin, "Esas üzücü olansa takipçilerimizin sonuna kadar haklı olmaları. Açıkcası bizim de içimize sinmedi yani. Kendi aracımız olmasa biz bile like vermezdik o fotoğraflara. İnsanın eli gitmiyor..." sözleriyle programın başarısız olduğunu kabul etti.

Harcanan onca para ve zamanın ardından gelen fotoğrafların bir Cardi B. makyajsız selfie'si kadar bile like alamadığına dikkat çeken Watzin, "Bir kaç gün içinde bir grup marslının çiftleşme fotoğrafı gibi bir şeyler gelmezse programın maliyetini çıkarması için gereken like sayısına ulaşmamız şu an için imkansız görünüyor. Resmen attığımız taş ürküttüğümüz kurbağaya değmemiş durumda. Neden böyle oldu? Kameraları mı düzgün seçmedik? Mars'ın kendisi mi fotojenik değil? Bunlar hep cevaplanması gereken sorular" derken, sorunun kaynağı anlaşılana kadar Mars programına ara verdiklerini açıkladı.

Figure 6.3: Human annotated full article C

A.6 Article C - Translation

NASA Announces Termination of Mars Program Due to Photos Not Reaching Expected Like Counts…

The American Space Agency NASA is facing troubled days due to the photos sent by the Perseverance rover following its landing on Mars.

NASA's one-ton Rover-type spacecraft, Perseverance, landed successfully on Mars' Jezero Crater at 15:55 Eastern Time on Thursday after a journey of approximately 7 months. However, 24 hours after the landing, when NASA officials shared the first photos from the red planet on their Instagram account, they expressed surprise and disappointment over the comments they received. The photos, which were expected to garner at least 10 million likes, only managed to accumulate 2 million, leading to significant disillusionment within the community.

Comments such as "Did you send an iron to Mars to take these photos?", "You'd think they'd put a decent camera on it", and "There was no need to go all the way to Mars for this, we could have sorted it out in Yozgat city center" were particularly hurtful, according to James Watzin, Director of NASA's Mars Program. Watzin admitted, "What's truly upsetting is that our followers are absolutely right. Honestly, we weren't satisfied either. If it wasn't our own rover, we wouldn't have liked those photos. It's just not appealing…"

Watzin pointed out that the photos failed to receive as many likes as a makeup-free selfie by Cardi B, noting, "If we don't receive something like a group of Martian mating photos in the next few days, it seems impossible to reach the number of likes needed to cover the program's costs. The effort we put in seems to have missed its mark. Why did this happen? Did we choose the wrong cameras? Is Mars itself not photogenic? These are questions that need answering." He announced that the Mars program will be put on hold until the source of the issue is understood.

B ChatGPT Explanations of Articles A, B, and C

B.1 Explanation by ChatGPT for Article A - Turkish

Bu haber metni de açıkça satirik bir içerik taşımaktadır. İşte nedenleriyle birlikte satirik kısımlar:

- **Başlık:** "LUNA'da Kriz Büyüyor: TL'ye Karşı Bile Değer Kaybettiği Ortaya Çıktı…"
  - **Neden:** Kripto paraların genel olarak değeri ABD doları karşısında değerlendirilirken, Türk Lirası'na (TL) karşı bile değer kaybetmesi abartılı ve mizahi bir ifade taşıyor.
- **Giriş:** "Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından..."
  - **Neden:** Bu ifade, kripto paraların değer kaybını dramatik bir şekilde betimleyerek abartı ve mizah katıyor.
- **Do Kwon'un Açıklaması:**
  - "Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken..."
    * **Neden:** TL'nin istikrarsız ve düşük değeri bilinirken, bir kripto paranın ona karşı bile değer kaybetmesi absürt bir durum olarak mizahi bir etki yaratıyor.
  - "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım."
    * **Neden:** CEO'nun TL'ye karşı değer kazanmanın bir moral kaynağı olabileceğini düşünmesi komik ve alaycı bir durum yaratıyor.
- **Ekip Reaksiyonu:**
  - "O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz."

    * **Neden:** Durumun ciddiyeti ile ilgili komik bir abartı içeriyor. Çalışanların ofisi terk etmesi, olayın ciddiyetinin mizahi bir şekilde abartılmasıdır.
  - "Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten."
    * **Neden:** Ofis eşyalarının satılabilecek duruma gelmesi, krizin büyüklüğünü komik ve abartılı bir şekilde betimliyor.

- **Geleceğe Yönelik Beklentiler:**
  - "Kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek..."
    * **Neden:** Türkiye'nin ekonomik kararlarının saçma olarak tanımlanması ve bu kararların TL'yi daha da değersiz hale getireceğinin ima edilmesi, mizahi ve alaycı bir ifade taşıyor.
  - "İleriye dönük iyimser mesajlar vermeyi de ihmal etmedi."
    * **Neden:** Durumun tamamen umutsuz olmasına rağmen iyimser mesajlar verilmesi, durumu daha da komik hale getiriyor.

Bu nedenlerle, metin genel olarak kripto para dünyasındaki ciddi bir değersizleşme olayını abartılı ve alaycı bir dille ele alarak satirik bir içerik oluşturuyor.

B.2 Explanation by ChatGPT for Article A - English

This news article is also clearly satirical in nature. Here are the satirical parts along with the reasons:

- **Headline:** *"Crisis in LUNA Deepens: Found to Be Losing Value Even Against TL. . . "*
  - **Reason:** The notion that LUNA is losing value against the Turkish Lira (which is generally perceived as weak) is absurd and creates a humorous situation.

- **Introduction:** *"After experiencing a 99.8% drop in value, plummeting from $64 to $0.2 in 2 days, bad news continues to come from Terra Luna Coin."*
  - **Reason:** Such a significant drop in value should be a serious issue, but the article treats the event with irony.
- **CEO's Statements:**
  - *"He stated that they found out LUNA was losing value not only against the US dollar but even against TL, saying, 'In the past two days, everything we said couldn't happen has happened. I'm very sorry.'"*
    - **Reason:** Presenting the loss of value against TL as unexpected and comical.
  - *"He noted that the situation emerged during a meeting held to boost the team's morale..."*
    - **Reason:** The irony of a morale-boosting meeting revealing an even more demoralizing situation creates humor.
  - *"'Look, it's not that bad. At least we are still gaining value in TL terms,' I gathered the team to make a motivational speech..."*
    - **Reason:** The CEO's attempt to lighten the situation with an ironic and humorous statement.
  - *"At that moment, half the team left the building. We still can't reach them."*
    - **Reason:** The exaggeration of team members leaving the building adds a humorous twist to the seriousness of the situation.
  - *"The remaining ones are staying just to see what can be sold in the office."*
    - **Reason:** The company's impending bankruptcy is expressed in an absurd and humorous way.
- **Forward-Looking Optimistic Messages:**
  - *"Expecting that in the short term, some nonsensical move by Turkey or the CBRT spending the last foreign exchange reserves might cause TL to depreciate more than LUNA, giving optimistic messages for the future..."*
    - **Reason:** Mocking Turkey's economic management by creating an absurd and humorous expectation.

For these reasons, the article overall treats a serious financial crisis in a humorous and ironic manner, creating a satirical content.

B.3 Explanation by ChatGPT for Article B - Turkish

Bu haber metni açıkça satirik bir içerik taşımaktadır. İşte nedenleriyle birlikte satirik kısımlar:

- **Başlık:** *"Uyarı: Will Smith'in Attığı Tokatla İlgili Görüş Bildirmeyen Bir Tek Siz Kaldınız..."*
  - **Neden:** Başlık, okuyucuya mizahi bir uyarı yaparak herkesin bu olay hakkında fikir beyan ettiğini ima ediyor. Bu, abartılı ve komik bir yaklaşım.
- **Giriş:** *"gözler bu konuda henüz görüşünü bildirmemiş son sosyal medya kullanıcısı olan size çevrildi."*
  - **Neden:** Bu ifade, olayın abartılı bir şekilde herkes tarafından yorumlandığını ve son görüşü bildirecek kişinin okuyucu olduğunu mizahi bir dille vurguluyor.
- **Bilgi maddeleri:**
  - *"Will Smith'in karısı kanser değilmiş. Saçkıran mı ne öyle dandik bir hastalık yüzünden saçını kazıtmış"*
    - **Neden:** Hastalığın önemsizleştirilmesi ve "dandik bir hastalık" ifadesi mizah ve alay içeriyor.
  - *"Evet hakkaten vurmuş. Ama yumruk değil tokat"*
    - **Neden:** Bu ifade, olayın şiddetini azaltarak mizahi bir dille veriyor.
  - *"Toksik maskülinite" kalıbını cümle içinde kullanırken dikkat edin. Yanlış yazan çok var."*
    - **Neden:** Bu, sosyal medyada sıkça yapılan dilbilgisi hatalarıyla dalga geçiyor.
  - *"Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı"*
    - **Neden:** Duygusal anın mizahi bir dille alaya alınması.
  - *"Olay kurgu değil. Ama ola dabilir. Ya da yok ya değil..."*

        * **Neden:** Olayın gerçekliğini sorgulayan ve çelişkili ifadelerle kafa karıştıran mizahi bir yaklaşım.

  - *"Tuvalet kağıdında KDV oranı %8'e indirildi (Belki bir faydası olur)"*

        * **Neden:** Alakasız bir bilgi eklenerek konunun ciddiyetinin düşürülmesi.

  - *"Chris Rock haklı. Bunda düşünecek bir şey yok"*

        * **Neden:** Basit ve kesin bir yargı ile olayın karmaşıklığının küçümsenmesi.

Bu nedenlerle, metin genel olarak olayın ciddiyetini mizahi ve alaycı bir dille ele alarak satirik bir içerik oluşturuyor.

B.4 Explanation by ChatGPT for Article B - English

This news article is clearly satirical in nature. Here are the satirical parts along with the reasons:

- **Title:** *"Warning: You Are the Last Person Who Hasn't Shared an Opinion About Will Smith's Slap..."*

  - **Reason:** The title humorously warns the reader that everyone has shared their opinion on this event, implying an exaggerated and comedic approach.

- **Introduction:** *"all eyes are now on you, the last social media user who hasn't shared their opinion on this matter."*

  - **Reason:** This statement humorously emphasizes that everyone has exaggeratedly commented on the event and the reader is the last person to share their opinion.

- **Information Points:**

  - *"Will Smith's wife doesn't have cancer. She shaved her head due to some trivial disease called alopecia."*

        - **Reason:** The trivialization of the disease and the phrase "trivial disease" contain humor and mockery.
- *"Yes, he really slapped him. But it was a slap, not a punch."*
    - **Reason:** This statement humorously downplays the severity of the incident.
- *"Be careful when using the term 'toxic masculinity' in a sentence. Many people spell it wrong."*
    - **Reason:** This mocks the frequent grammatical errors on social media.
- *"After slapping, Will Smith went and cried his way to an Oscar."*
    - **Reason:** The emotional moment is humorously ridiculed.
- *"The event is not staged. But it could be. Or no, it's not..."*
    - **Reason:** The humorous approach of questioning and contradicting the reality of the event.
- *"The VAT rate on toilet paper has been reduced to 8% (Maybe it will help)."*
    - **Reason:** Adding irrelevant information to diminish the seriousness of the topic.
- *"Chris Rock is right. There's nothing to think about here."*
    - **Reason:** Simplifying and trivializing the complexity of the event with a straightforward judgment.

For these reasons, the article overall treats the seriousness of the event with humor and ridicule, creating satirical content.

B.5 Explanation by ChatGPT for Article C - Turkish

Bu haber metni de satirik bir içerik taşımaktadır. İşte nedenleriyle birlikte satirik kısımlar:

- **Başlık:** *"NASA, Fotoğrafların Beklenen Like Sayısına Ulaşamaması Nedeniyle Mars Programını Sonlandırdığını Açıkladı..."*
  - **Neden:** NASA gibi bilimsel bir kurumun sosyal medya beğenileri nedeniyle önemli bir programı sonlandırması absürt bir durumdur. Bu, okuyucuda mizahi bir etki yaratır.
- **Giriş:** *"Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor."*
  - **Neden:** Bilimsel bir başarıyı sosyal medya beğenileriyle ilişkilendirmek absürt ve komiktir.
- **Fotoğrafların Beklenen Beğeni Sayısına Ulaşamaması:**
  - *"En az 10 milyon beğeni alması beklenen fotoğrafların 2 milyonda kalması da camiada büyük hayalkırıklığına neden oldu."*
    - **Neden:** Mars keşfi gibi ciddi bir bilimsel olayın sosyal medya beğenileriyle ölçülmesi saçma ve mizahi bir yaklaşımdır.
- **Yorumlar ve Tepkiler:**
  - *"Mars'a ütü yolladınız da o mu çekti fotoğrafları?", "İnsan şuna bi tane düzgün kamera koyar", "Bunu çekmek için Mars'a kadar gitmeye gerek yoktu, Yozgat şehir merkezinde de hallederdik"*
    - **Neden:** Bu yorumlar, olayın ciddiyetini alaycı bir şekilde küçümseyerek mizah yaratır.
  - *NASA yetkilisinin bu yorumları haklı bulması ve "Kendi aracımız olmasa biz bile like vermezdik o fotoğraflara."*
    - **Neden:** NASA yetkilisinin bu kadar önemli bir projeyi sosyal medya beğenilerine göre değerlendirmesi ve kabul etmesi absürt ve komiktir.

- **Sonuç ve Açıklamalar:**
  - *"Harcanan onca para ve zamanın ardından gelen fotoğrafların bir Cardi B. makyajsız selfie'si kadar bile like alamadığına dikkat çeken Watzin..."*
    * **Neden:** Bilimsel bir projenin sosyal medya fenomenleriyle kıyaslanması gülünç bir durum yaratır.
  - *"Bir kaç gün içinde bir grup marslının çiftleşme fotoğrafi gibi bir şeyler gelmezse programın maliyetini çıkarması için gereken like sayısına ulaşmamız şu an için imkansız görünüyor."*
    * **Neden:** Marslıların çiftleşme fotoğrafları gibi absürt bir beklenti ve sosyal medya beğenilerinin maliyet karşılaştırması komik ve alaycı bir yaklaşım.
  - *"Resmen attığımız taş ürküttüğümüz kurbağaya değmemiş durumda. Neden böyle oldu? Kameraları mı düzgün seçmedik? Mars'ın kendisi mi fotojenik değil?"*
    * **Neden:** Mars gibi ciddi bir bilimsel hedefin fotojeniklikle alaycı bir şekilde sorgulanması mizahi bir etki yaratır.

Bu nedenlerle, metin genel olarak ciddi bir bilimsel çalışmanın önemini sosyal medya beğenileriyle alay ederek küçümseyen ve mizahi bir dille ele alan satirik bir içerik oluşturuyor.

B.6 Explanation by ChatGPT for Article C - English

This news article is also clearly satirical in nature. Here are the satirical parts along with the reasons:

- **Title:** *"NASA Announces It Is Ending Mars Program Due to Photos Not Getting Expected Number of Likes..."*
  - **Reason:** The absurdity of a scientific institution like NASA ending an important program due to social media likes creates a humorous effect for the reader.

- **Introduction:** *"The American Space Agency NASA is having a hard time due to the photos sent by the Perseverance rover after its landing on Mars."*
  - **Reason:** Associating a scientific achievement with social media likes is absurd and comical.
- **Photos Not Reaching Expected Number of Likes:**
  - *"The photos that were expected to get at least 10 million likes ended up with only 2 million, causing significant disappointment in the community."*
    - **Reason:** Measuring a serious scientific event like Mars exploration with social media likes is a ridiculous and humorous approach.
- **Comments and Reactions:**
  - *"You sent an iron to Mars, and that took the photos?", "One could put a decent camera on that", "There was no need to go all the way to Mars to take these pictures; we could have done it in Yozgat city center."*
    - **Reason:** These comments create humor by sarcastically belittling the seriousness of the event.
  - *The NASA official finding these comments valid and saying, "If it weren't our own vehicle, we wouldn't even like those photos."*
    - **Reason:** It is absurd and comical for a NASA official to evaluate such an important project based on social media likes and to accept this evaluation.
- **Conclusions and Statements:**
  - *"Watzin pointed out that the photos received fewer likes than a makeup-free selfie of Cardi B..."*
    - **Reason:** Comparing a scientific project to social media phenomena creates a ridiculous situation.
  - *"If something like a group of Martians mating photos doesn't come in a few days, it currently seems impossible to reach the number of likes needed to cover the cost of the program."*

      * **Reason:** The absurd expectation of Martian mating photos and the comparison of social media likes to program costs is a comical and sarcastic approach.

- *"The stone we threw hasn't hit the frog we scared. Why did this happen? Did we choose the cameras incorrectly? Is Mars itself not photogenic?"*
      * **Reason:** The sarcastic questioning of Mars, a serious scientific target, based on photogenic qualities creates a humorous effect.

For these reasons, the text overall forms a satirical content that mocks and diminishes the importance of a serious scientific endeavor by making fun of social media likes.

C Other Articles

C.1 Article D

Bu sabah Taraf gazetesi tarafından ortaya atılan "Tutsaklara karşılık Süleyman Şah Türbesi IŞİD'a verilecek" şeklindeki şok haber, Dışişleri Bakanlığı tarafından kesin bir dille yalanlandı. Önce internet sitesinden yapılan açıklama sonra da Basın Sözcüsü Kamuran Aydilen aracılığı ile kamuoyunu aydınlatan Dışişleri Bakanlığı, "Ortadaki yanlış anlaşılmaları gidermek için söylüyoruz, Süleyman Şah Türbesi'nin yıkılarak yerine AVM yapılması konusunda IŞİD'le görüştüğümüz doğru. Neticede türbe yıkımında kendilerinden daha tecrübeli bir ekip yok. Ancak bunun dışında herhangi bir pazarlık söz konusu değil" ifadeleri ile iddiaları reddetti.

Bakanlık binasında gazetecilerin sorularını yanıtlayan Bakanlık Sözcüsü Aydilen, türbenin yıkım ihalesi için IŞİD ile pazarlık masasında oturulduğunu itiraf ederken, konunun rehinlerle doğrudan bir ilgisi bulunmadığını ise şu sözlerle savundu:

"Arkadaşlar 12 yıllık iktidarımızda artık bizi biraz tanımış olmanız lazım. Bütün dünya bilir ki biz, öyle 49 kişi için bir karış toprak vermeyiz. Hele de öyle bir toprağı, tam kupon arazi orası, deli misiniz ya? Mümkün mü böyle bir pazarlık? Türbeyi de geç, sırf arsası 4 milyar dolar eder. Orada nöbet tutan askerlerimize de sorduk, çevrede başka AVM de yokmuş. 'Çarşı izninde gidecek yer bulamıyoruz' diyorlar.

Şu inşaat bir başlasın, Allah'ın izniyle para basacak orası..."

IŞİD'ın özellikle türbe yıkım işinde uzmanlaşmış, işlerini severek yapan ve sahiplenen bir örgüt olduğunun altını çizen Basın Sözcüsü, "Şu an bizden haber bekliyorlar, tamam dediğiimiz anda havanlarla falan girişecekler. Alimallah 1 saatte taş üstüne taş koymayız dediler. Rehineler konusunu öyle özel olarak konuşmadık ama o konuda bir jest yaparlarsa biz bunu geri çevirmeyiz elbette. Neticede birlikte iş yapan insanlarız, yarın öbür gün başka yıkım ihaleleri de olur... Bunları da değerlendireceklerdir" ifadelerine yer verdi.

Mevcut anlaşmanın devletin kasasından bir kuruş çıkmadan halledileceğinin üzerinde duran Aydilen, yapılması planlanan AVM'nin detaylarını da basın mensuplarıyla paylaştı:

"Bakın buradan bööyle şimdiki türbenin kubbesi şeklinde bir tavan geliyor. Orası food court olacak... Alt katta SHAH'S SPORT adında bir fitness salonu ve atış poligonu var. Ta buraya kadar da meydan, forum mantığı gibi düşünün siz. Şimdi tabii aklınıza hemen ulaşım işi geliyor... Onu da düşündük. Hızlı treni 2017'de Marmaray'la Halkalı'ya bağladıktan sonra, Halkalı Ankara arası 4.5 saate inmiş olacak. Ankardan da ring seferiyle tak Halep'tesin. Son olarak Halep - Karakoza arası İDO'nun motorlarına binecek vatandaşlarımız anında AVM'de olacak. Bu kadar basit. Ayrıca oradaki askerlerimizi de özel güvenlik ve otopark görevlisi olarak AVM'de istihdam etmeyi düşünüyoruz. Gördüğünüz gibi bu projede kaybeden yok..."

Bir soru üzerine Suleymanium AVM'yi, yaşasaydı Suleyman Şah'ın da takdirle karşılayacağını sözlerine ekleyen Dışişleri Sözcüsü, son olarak şunları kaydetti:

"Yani düşünün tarihe geçmiş bir şahısınız, arkanızda bir tanecik kullanılmayan türbe kalıyor. Ne sineması var, ne otoparkı... Böyle mi anmalıyız ecdadımızı? Ayrıca son dönemde biliyorsunuz TOKİ'nin mevcut tarihi yapılar etrafında çeşitli çalışmaları mevcut. Sosyal medyada tarihi kümbetle iç içe geçmiş yurtlarımız büyük ilgi gördü. Bu şekilde alışveriş keyfini manevi iklimle birleştiren bir çalışma halkımızın da ilgisini çekecektir..."