TU WIEN Informatics

Explainable Prediction of User Post Popularity: An Analysis of the One Million Posts Corpus

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Dario Bogenreiter, MSc
Matrikelnummer 11702132

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass. Gábor Recski, PhD
Mitwirkung:

Wien, 1. Jänner 2001

Dario Bogenreiter	Gábor Recski

TU WIEN Informatics

Explainable Prediction of User Post Popularity: An Analysis of the One Million Posts Corpus

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Dario Bogenreiter, MSc
Registration Number 11702132

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. Gábor Recski, PhD
Assistance:

Vienna, January 1, 2001

Dario Bogenreiter

Gábor Recski

Erklärung zur Verfassung der Arbeit

Dario Bogenreiter, MSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. Jänner 2001

Dario Bogenreiter

Danksagung

Ihr Text hier.

Acknowledgements

Enter your text here.

Kurzfassung

Ihr Text hier.

Abstract

Enter your text here.

In the contemporary digital communication landscape, the influence wielded by news agencies, users, and even bots is undeniable. Their ability to shape public opinion, drive agendas, and potentially sway elections has sparked interest in understanding the factors governing post popularity. While previous research on post popularity prediction has predominantly focused on platforms like Twitter, this thesis ventures into uncharted territory by analyzing the One Million Posts Corpus, sourced from the Austrian daily newspaper *Der Standard*. By delving into this dataset, which represents a unique demographic and content context, this study aims to unravel the details of post-engagement dynamics and showcase the underlying mechanisms that lead to popularity trends.

Contents

CHAPTER 1

Introduction

1.1 Motivation & Problem Statement

Even before the 2017 US presidential election between Trump and Clinton, it was evident that user posts on social media and news websites had a significant influence. They can shape opinions, drive agendas, and potentially sway elections. The visibility of these posts often depends on the number of likes and dislikes they receive, with highly liked posts being prioritized. Understanding the reasons why certain posts trend and attract varying numbers of likes is crucial.

Social media popularity prediction (SMPD) is the endeavor to predict the popularity of content posted on social networks [DWW19]. In the past, research within this area focused mainly on data from Twitter [DGVM20] and other standard social media [LWX+22, LZZ20]. However, when discussing and forming their political opinions, online communities of newspaper websites play a major role [BM12]. Such websites are not classic social media websites. Nevertheless, since they allow users to create posts, share personal stories, and interact with each other, these websites share major characteristics of standard social media websites. Although their influence on the political discussion can not be neglected, only a handful of papers like [RK20a] have explored predictive models for these kinds of websites.

There exists a research gap in the explainable prediction of user post popularity on news websites. Current research lacks detailed, explainable analyses of why certain features and algorithmic solutions work. Qualitative analysis is crucial to unveil the mechanism of why certain posts are trending and others are not. Understanding the qualitative aspects of these predictions is vital for interpreting why certain content resonates with users. Additionally, most studies have focused on English texts and news, neglecting German use cases that may have different characteristics.

The Problem

While certain characteristics influencing the popularity of posts on social media and news platforms overlap across different online communities, each community exhibits unique features and dynamics. These differences can arise from various sources such as the user base or content ranking mechanisms. For instance, posts on the website of the British newspaper *The Guardian* are ranked strictly in chronological order, whereas posts on the Austrian newspaper *Der Standard* are ranked based on their popularity. Such variations underscore the need to consider community-specific factors when analyzing post popularity. For this reason, this research specifically focuses on understanding and predicting post popularity specifically for *Der Standard*. As one of the primary platforms for political discussions in Austria, it plays a crucial role in shaping public opinion and influencing political discourse. Analyzing the dynamics of this community is essential, as the discussions on this platform can significantly impact political outcomes, including elections.

The most appropriate dataset for this research is the 'One Million Posts Corpus' [SST17], which contains over one million user posts from *Der Standard*, along with their upvotes and downvotes. This dataset offers rich training and testing data for building models to predict the popularity of user-generated content. However, so far, research on this dataset has been limited to tasks such as detecting offensive posts [SST17, WRJB18], leaving a gap in understanding the factors driving post popularity and the role of explainability in these predictions.

1.2 Research Questions

The research is structured around the following core questions:

R1. What are the most suitable explainable features for predicting post popularity?

R2. How do different state-of-the-art, black-box models, such as BERT or RoBERTa, perform in predicting post popularity compared to simple baseline models and the explainable models introduced in this work?

- R2.1. Do the same models that perform best in classification tasks also perform best in regression tasks?
- R2.2. Does the threshold for identifying top comments in the classification task influence the results? Do the results vary for different classes of high-engagement posts?
- R2.3. Do any of the models work best for certain subsets of data (e.g., posts with high versus low engagement)?

These research questions correspond to the following null hypothesis:

$H_0$1. There is no significant difference in feature importance.

$H_0$2. There is no significant performance difference between state-of-the-art black-box models (such as BERT and RoBERTa), simple baseline models, and the explainable models introduced in this work when predicting post popularity.

- $H_0$2.1. Models that perform best in classification tasks do not perform significantly better or worse in regression tasks.
- $H_0$2.2. The choice of threshold for identifying top comments in the classification task does not significantly influence the results. There is no significant variation in results for different classes of high-engagement posts.
- $H_0$2.3. None of the models show significantly better performance for specific subsets.

1.3 Expected Contribution

This thesis aims to develop a structured pipeline for creating explainable features to predict post popularity, introduce a taxonomy to identify top and engaging comments, and provide a detailed evaluation of the algorithmic solutions applied. The code base for this project will be made publicly available in a GitHub repository[1] under the MIT license[2].

Rather than solely prioritizing quantitative prediction accuracy, this research focuses on understanding the underlying factors that drive user engagement and seeks to connect these findings with prior research. By addressing the interplay between algorithmic performance and explainability, this work intends to contribute a comprehensive framework for understanding and predicting user engagement patterns.

The answers to these research questions will form the foundation for a comprehensive evaluation of the prediction models and contribute to a structured understanding of the mechanisms influencing post popularity and user engagement.

1.4 Structure of the thesis

Section 2 reviews related work, covering both studies that utilize the same dataset and research within the broader area. Section 3 introduces the dataset, highlights its unique characteristics, and provides a detailed data analysis. Section 4 outlines the research methods and experimental setup. The results are presented in Section 5, followed by an in-depth discussion and analysis in Section 6. Finally, Section 7 concludes the thesis by summarizing key findings, discussing the limitations, and suggesting directions for future research.

[1] https://github.com/dario-x/user_post_popularity_prediction_DerStandard
[2] https://opensource.org/licenses/MIT

CHAPTER 2

Related Work

As previously mentioned, the dataset used in this research was originally designed for text classification tasks, mainly for sentiment analysis and hate speech detection [SST17]. The authors of the dataset developed several approaches to predict whether a post falls into one of nine predefined categories: *Negative Sentiment*, *Positive Sentiment*, *Neutral Sentiment*, *Off Topic*, *Inappropriate*, *Discriminatory*, *Asking for Feedback*, *Shares Personal Story*, and *Uses Rational Argumentation and Reasoning* [SST17].

For baseline solutions, they employed a Bag of Words (BOW) model in conjunction with a Support Vector Machine (SVM). Additionally, they explored more sophisticated solutions, such as a doc2vec (D2V) document embedding combined with an SVM, and a neural network architecture using a Long Short-Term Memory (LSTM) model [SST17]. The performance of these models was evaluated using precision, recall, and F1 score [SST17]. Interestingly, the performance varied across the models and categories [SST17]. The BOW + SVM model provided robust results across most tasks, achieving the highest precision in two cases. In contrast, the more advanced LSTM model outperformed others in four task-metric combinations but notably failed to identify any posts with positive sentiment [SST17]. This highlights the complexity of text classification tasks and the varying effectiveness of different models depending on the specific label and evaluation metric.

In 2020, Scheible et al. [STT+20] utilized the One Million Posts dataset to evaluate their newly developed transformer model, GottBERT, based on the classification task, discussed in the previous paragraph. At the time, GottBERT was the first RoBERTa [LOG+19] model specifically designed as a monolingual model for German [STT+20]. Despite being trained on 145 GB of data from sources such as Wikipedia, the EU Bookshop corpus, and the German portion of the Open Super-large Crawled ALMAnaCH Corpus (OSCAR), GottBERT did not meet expectations [STT+20]. It was outperformed in terms of F1 score by other BERT models and even by the multilingual XLM-RoBERTa, which was

unexpected given that multilingual models typically perform worse than monolingual transformer models [STT+20]. The authors suggested that this underperformance could be attributed to sub-optimal hyperparameters used during the model's training [STT+20]. This is a logical suggestion, as the researchers did not conduct solid hyperparameter testing for this study.

In 2018, Wiedemann et al. [WRJB18] also focused on the 11,773 labeled posts of the dataset, utilizing this data as background knowledge for transfer learning. They conducted supervised pre-training of a neural model on the task of classifying offensive content [WRJB18]. However, this approach resulted in only minor improvements [WRJB18]. A possible explanation for the underwhelming results could be that they used the labels "inappropriate" and "discriminatory" to approximate whether a comment was "offensive" or not [WRJB18]. While inappropriateness and discrimination are closely related to offensiveness, according to the authors, offensiveness cannot be fully described by these two labels [WRJB18]. Consequently, they referred to this approach as near-category transfer learning [WRJB18].

Risch and Krestel emphasized the value of the labeled part of the dataset [RK20b], who listed it as a common dataset for supervised training on the detection of toxic comments. They defined toxic comments as a complex concept primarily employed by spammers, haters, and trolls, which reduces user engagement on the platform [RK20b]. Furthermore, they explained that toxicity is a challenging topic because users who post toxic comments often intentionally try to conceal the actual meaning of their posts. Stylistic devices such as irony further hinder classification [RK20b]. These remarks are valuable as these challenges could also apply when determining the popularity of a comment. For instance, some top comments might exhibit a high level of irony and be celebrated for this circumstance, as they are perceived as particularly funny.

In 2020, Risch and Krestel studied predicting the number of upvotes and replies to posts on the newspaper *The Guardian* [RK20a]. This research is closely related to our study but differs in three ways: they did not consider downvotes, applied different algorithmic approaches, and the scenario is entirely different (English texts with posts ranked chronologically rather than by votes received). Their primary motivation was to demonstrate an automated method for identifying engaging posts without expensive manual annotation effords by editors [RK20a]. Such a method could improve user experience by recommending posts for readers to read or reply to [RK20a].

They defined popularity by the number of upvotes a post received and engagement as a combination of the adjusted number of upvotes and replies [RK20a]. This definition was justified by the assumption that users upvote posts that interest them the most and that users do not manipulate votes, as upvoting does not affect post ranking (posts are chronologically sorted on *The Guardian*) [RK20a]. This assumption does not hold for this research, as more upvotes on a post on *Der Standard*'s website result in a higher ranking. Their target variable, "top" and "flop" posts, were defined based on the relative

share of upvotes each post received compared to all posts under one article [RK20a].

For predicting "top" and "flop" comments, they employed four approaches: a baseline logistic regression trained on text length, logistic regression with features proposed by Park [PSDE16], a convolutional neural network (CNN), and their newly designed recurrent neural network [RK20a]. The features of Park [PSDE16] included text length, readability, and averages of text length, number of received comment upvotes, and readability per user. Their newly proposed solution outperformed the other models by a few to a maximum of 10 percent in accuracy, depending on the percentage of comments identified as "top comments" [RK20a].

Most interestingly, they identified challenges similar to those with the 'One Million' dataset, such as earlier comments receiving more upvotes, which they referred to as position bias, as comments are ranked chronologically [RK20a]. They corrected for this bias by grouping comments into ranks according to the time they were written and then calculating "top" and "flop" relative to these ranks [RK20a]. Another challenge they faced, which also exists in 'One Million' dataset, is that some articles have very few comments, making it difficult to correctly estimate popularity. They addressed this by removing such posts from their analysis [RK20a].

Besides the endeavor of identifying valuable and engaging user posts by the number of votes they receive, there has been ongoing research on how articles could be classified as particularly valuable by looking at the posts (comments) that they receive [TWDR09, ARKL18, BAH12]. Tsagkias et al. applied Random Forests to identify popular articles through a two-step process [TWDR09]. First, they determined if users would comment on an article at all. Second, they predicted the number of comments that would appear under the article [TWDR09]. They noted that the second task is significantly more challenging, and accurately predicting the exact number of comments is practically infeasible [TWDR09]. Similarly, Bandari et al. studied this problem in the context of social media and reached comparable conclusions [BAH12]. They found that predicting popularity as a regression task results in large errors, so they redefined it as a classification task by grouping articles based on the number of comments they receive [BAH12]. These findings could also be relevant to this research, as the dynamics of the posts under an article might have similar dynamics to the articles themselves.

CHAPTER 3

The 'One Million Posts Corpus' dataset

The One Million Posts Corpus dataset [SST17] contains information on over one million comments related to articles from the Austrian daily newspaper *Der Standard*. This dataset, created by the Austrian Research Institute for Artificial Intelligence (OFAI), originates from the newspaper's website, where registered users can post comments below news articles [SST17]. Users can also reply to earlier comments, creating tree-like discussion thread structures [SST17].

The dataset includes the following data columns for these posts:

- ***Post ID*** - unique identifier for each post
- ***Article ID*** - identifier for the article in whose comment section the post appears
- ***User ID*** - anonymized identifier for the user who commented
- ***Headline*** - headline of the post (max. 250 characters)
- ***Main Body*** - main content of the post (max. 750 characters)
- ***Time Stamp*** - time when the post was created
- ***Parent Post*** - identifier for the parent post if the comment is a reply
- ***Status*** - indicates if the post is online or was deleted by a moderator
- ***Positive Votes*** - number of positive votes by other users
- ***Negative Votes*** - number of negative votes by other users

Additionally, the dataset includes details about the articles under which users have posted their comments. This information includes:

- ***Article ID*** - unique identifier for each article
- ***Path*** - topic of the article (e.g., 'Newsroom/Sports/')
- ***Publishing Date*** - timestamp of when the article was published
- ***Title*** - headline of the article
- ***Body*** - full text of the article

The dataset contains 11,773 labeled posts and an additional one million unlabeled posts (totaling 1,011,773 posts) [SST17]. "Labeled" refers to posts categorized into seven relevant categories for hate speech detection and sentiment analysis. This dataset was originally designed to explore the task of moderating discussions in online forums, utilizing both machine-learning tools and professional human moderators [SST17]. However, the authors noted that the dataset could support additional use cases [SST17]. One such use case is predicting post popularity based on positive and negative vote counts. This task benefits from having an indisputable ground truth for over one million posts, unlike sentiment detection, which often suffers from subjectivity issues.

Out of the 1.01 million posts, 0.69 million have received at least one upvote or downvote, as shown in Table 3.1. Specifically, 0.63 million posts have received at least one upvote, while only 0.3 million posts have received at least one downvote. Overall, upvotes are more common than downvotes. The median (1) and mean (3.78) number of upvotes per post are significantly higher than the median (0) and mean (1.08) number of downvotes per post. This disparity may result from several factors: offensive comments being deleted, users being more reluctant to post comments that might receive negative attention, or the tendency of like-minded individuals to comment on certain articles. Further research is needed to determine the exact reasons.

Table 3.1: Number of Posts in Different Categories

Category	**Number of Posts (in millions)**
All Posts	1.01
Posts with Votes	0.69
Posts with UpVotes	0.63
Posts with DownVotes	0.30

Figure 3.1 shows a heatmap of the posts, clustered according to the number of upvotes and downvotes. Each cell represents the count of posts with a specific number of upvotes and downvotes. For example, 126 posts have 0 downvotes and between 100 and 500 upvotes. Summing up all the cells in the heatmap gives 1.01 million posts. The heatmap shows that most posts with more than 100 upvotes have either 0 or just a few downvotes. Interestingly, all posts with more than 100 downvotes have at least one upvote. In addition, we can see that about 0.32 million posts have 0 up- and downvotes, which is the highest count for any combination of up- and downvotes.

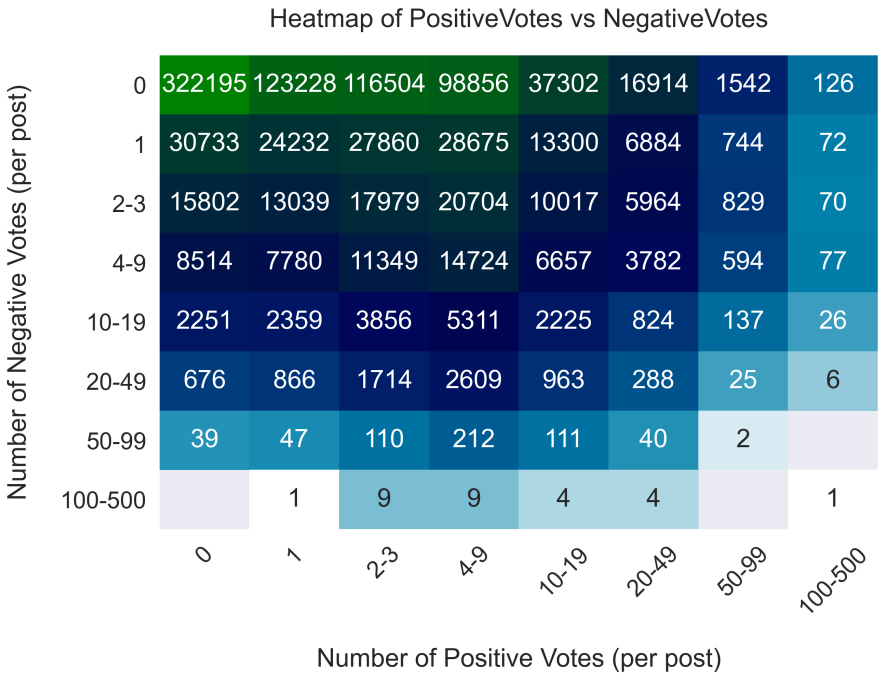

Figure 3.1: Heatmap of Posts, clustered by Up- and Downvotes

A reason why many posts receive very little attention (in terms of upvotes and downvotes) is simply that the vast number of posts limits what users can read [RK20a]. Consequently, many posts don't get enough feedback to determine whether they're perceived positively, neutrally, or negatively. Posts that are perceived as neutral present a general problem, as users do not have the option to give a post a neutral vote, and the number of reads per post is not collected.

A further limitation of the amount of attention a post can receive lies in the continuous publication of news articles, leading to older articles and their comments being forgotten. This naturally caps the amount of interaction a post can receive, to about 500 upvotes and 500 downvotes.

The Privilege of Top Comments

One reason why only few posts receive over 50 votes is that highly engaging posts are highlighted on the *Der Standard* website, as shown in Figure 3.2. These comments are listed above the text input field and all other posts. They remain in this prominent position until a post receives even more attention.

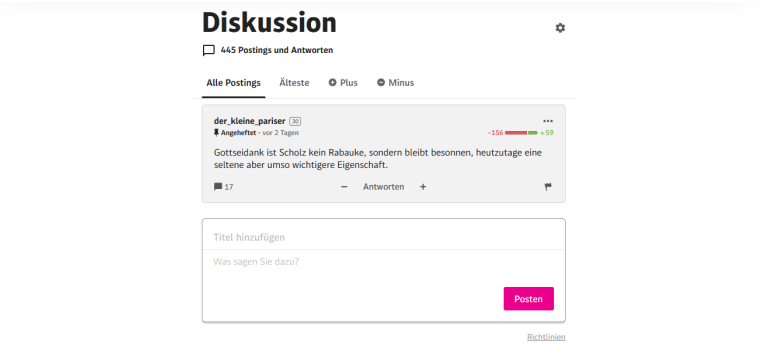

Figure 3.2: Example of a Top Comment

This layout, shown in Figure 3.2, boosts the visibility of popular posts, causing them to accumulate more votes while less prominent posts receive even less attention. This is why most other platforms have reverted to listing posts solely in chronological order [RK20a]. However, this approach overlooks the reality that some posts are inherently more engaging and deserving of extra attention [RK20a]. Additionally, it exacerbates the issue that the average user can only interact with a limited number of posts compared to the vast number available.

Result Distortion through Moderators

A possible explanation why upvotes are more common than downvotes might come from the fact that many posts contain offensive content or violate community guidelines and have therefore been removed - a problem, that has even led to newspapers shutting down their comments section (to save money on the employment of moderators checking these guidelines) [NKS21]. This naturally distorts the number of received up- and downvotes of these posts, as users can only see them for a limited time.

Bias of the dataset

The dataset contains data from a specific group of users. About 46 percent of all *Standard* readers have obtained an academic degree and the newspaper is the most read newspaper managers [1]. 39 percent of all readers are women. The average *Standard* reader is 47 years old. Moreover, the newspaper is recognized for its left-leaning political stance which naturally attracts a certain audience.

[1] https://www.derstandard.at/story/3000000212511/standard-ist-meistgenutztes-qualitaetsmedium-bei-entscheidungstraegern

CHAPTER 4

Features for Popularity Prediction

4.1 Literature Review

Predicting the popularity of online user-generated content, particularly posts involving upvotes and downvotes, has attracted significant attention in recent years due to the increasing use of social media and other content platforms. To ensure a comprehensive understanding of the various approaches and methodologies used to predict content popularity, a semi-structured literature review was conducted. This review largely follows the systematic literature review approach outlined by Kitchenham [Kit04], which provides a rigorous and reproducible method for identifying, evaluating, and synthesizing relevant studies in a field of interest. In the context of this study, the field of interest involves identifying features generated from user posts that can be utilized to predict their popularity.

The literature gathering process was divided into two strategies.

4.1.1 Gathering Stage

Gathering Method 1: Identifying Dataset-Specific Studies

In the first phase, Google Scholar was used to identify all papers that cited the original paper introducing the dataset utilized in this study. A total of 66 papers were collected in this manner.

Gathering Method 2: Broader Literature Search

In the second phase, a broader search was conducted to explore general research on predicting content popularity. For this, the ProQuest database was selected due to its ability to facilitate precise search queries and systematic filtering of results. The following search query was used to find relevant studies:

```
TIAB(("posts" OR "postings" OR "text" OR "textual") AND
("news" OR "reddit" OR "twitter" OR "facebook") AND
("prediction" OR "predict" OR "predictive" OR "machine learning")
AND ("popularity" OR "classification" OR
"detection" OR "sentiment" OR "polarity") AND
"features")
```

The keyword `TIAB` was used to limit the search to the title and abstract, ensuring that only papers directly relevant to the topic were gathered. The search query was constructed in several stages. First, textual data, including synonyms such as "postings" and "texts," was targeted. Next, the environment in which the posts occur, ranging from news websites to popular social media platforms, was captured to expand the potential field of investigation. Although news websites represent a smaller field, they may still provide inspiration for transferable features. Predictive tasks were then prioritized, including not only popularity prediction but also related fields such as sentiment analysis and hate speech detection, to yield potentially useful feature sets. Finally, the inclusion of "features" in the query ensured the collection of papers directly relevant to the extraction and use of features in prediction tasks.

By September 2024, this search query returned a total of 317 papers. These were subsequently filtered according to the inclusion criteria outlined below.

4.1.2 Inclusion Criteria & Analysis

The inclusion criteria for selecting papers were:

- Peer-reviewed scholarly journals to ensure a high standard of quality.
- Written in English.
- Published between January 2014 (2014/01/01) and September 2024 (2024/09/01) to ensure recent, up-to-date research (and as anyways more than 95 percent of the found studies were written in this time)

After applying these criteria, the search results were narrowed down to 173 papers, largely due to since 130 working papers from the original query results were excluded.

In total, 247 papers (66 from Gathering Method 1 and 181 from Gathering Method 2) were further analyzed based on their titles and abstracts to assess their relevance to the topic. Papers that explicitly listed the features used for prediction and had a least a remotely similar use case, as for example predicting the number of retweets or generating - where features that could be potentially be recycled could appear - where used for further analysis. Furthermore, this study focused on finding the explainable features, more complex features such as embeddings, are only briefly mentioned, as this they are not the core focus of this thesis.

4.1.3 Findings and Synthesis

In the final step, the information from all the papers was analyzed and then summarized in the following overview, which groups all named features and lists those papers that mentioned the respective feature. The features are grouped in the following categories: Pragmatic, Semantic, Lexical, Syntactic, and Meta.

Pragmatic Features

These features deal with the contextual and practical usage of language. They convey for example information on the tone and urgency.

- **Punctuation & Special Character Counts:** such as the frequency of punctuation marks or special characters, such as periods, exclamation points, brackets, hashtags, or question marks (for example to identify the number of questions as done in [HLM18]).
  Cited by 11 papers: [HLM18, EKHW23, JGM24, MCCH22, GS23, SPR22, ASM22, CKB+21, PLLGMP21, NPPZ18, BW15]

- **Capitalization:** The usage of lowercase or uppercase letters, which may be employed for emphasis (e.g., "THIS IS COMPLETELY WRONG!"). Other examples are the ratio between upper and lowercase letters or the count of fully capitalized words.
  Cited by 3 papers: [HLM18, EKHW23, ASM22]

- **Emojis Usage:** Measured for example in the frequency or proportion of emojis in a text relative to standard characters. Emojis may be classified as positive or negative and be useful in identifying the sentiment of a text [GIMW11, SASAW20].
  Cited by 6 papers: [EKHW23, SPR22, ASM22, CKB+21, SASAW20, GIMW11]

- **Text patterns:** The use of custom regular expressions to identify specific patterns that may convey particular meanings. For example, the presence of long words (longer than n characters) and short words (fewer than 4 characters) can be analyzed, as shown in [JGM24]. Another example would be to check whether a text starts with a letter, or a number [MCCH22]. Combinations such as "!!" may signify a certain tone that is associated with clickbait, as presented in [GS23, CPKG16].
  Cited by 4 papers: [HLM18, JGM24, MCCH22, GS23]

Lexical Features

These features relate to vocabulary and word usage in the text.

- **Text length:** measured, for example, in total or unique word count in the post or the number of characters. Additionally, counting sentences or syllables (using dictionaries such as LIWC) is possible, as suggested in [JGM24]. Metrics may include stop words or exclude them, as done in [PLLGMP21]
  Cited by 10 papers: [RK20a, KWR+24, JGM24, MCCH22, AHB+23, MS23, SPR22, CKB+21, PLLGMP21, SCM+17]

- **Bag of Words (BOW):** This feature counts occurrences of individual words, or unigrams, in a text.
  Cited by 9 papers: [ARKL18, RL21, KWR+24, MKR+24, DS24, KASA24, SD22, CKB+21, AA19]

- **Term frequency–inverse document frequency (TF-IDF):** This measure evaluates word importance relative to other documents. Words that appear frequently in a post but infrequently across all posts receive a high score.
  Cited by 19 papers: [WRJB18, ANM+21, RL21, HLM18, KA24, KWR+24, MKR+24, DS24, KASA24, T+23, AGAR+22, ASM22, SD22, GKSJ21, CKB+21, CKB+21, CKB+21, SASAW20, MW20]

- **N-grams:** N-grams are combinations of 2 words (bigrams), 3 words (trigrams), or any number of words (n). These words often appear together in specific contexts, providing meaningful insights.
  Cited by 12 papers: [ARKL18, ANM+21, RL21, KA24, DS24, KASA24, AGAR+22, CKB+21, SASAW20, MW20, AA19, BW15]

- **Lexical Diversity:** can be measured, for example by the number of unique words relative to the total text, as well as the counts of content and function words [JGM24].
  Cited by 10 papers: [JGM24]

- **Toxicity and Explicity Level:** This feature quantifies the usage of negative or swear words in a text, sometimes referred to as a profanity counter [SPR22]. Such a counter may also include explicit sexual language [SGS21].
  Cited by 5 papers: [ARD24, JGM24, SPR22, SGS21, SGS21]

- **Consistency:** measures how similar the title of a post is to its content, as demonstrated in [MCCH22]. It can also refer to whether the text follows a consistent theme by repeating certain words.
  Cited by 1 paper: [MCCH22]

- **Stopwords** This feature assesses the number of irrelevant terms or filler words in a text, which may be calculated as a ratio as well.
  Cited by 2 papers: [JGM24, SGS21]

Syntactic Features

These features relate to sentence structure and grammatical composition.

- **Part of Speech (POS) Tagging:** his technique identifies the grammatical roles of words and can be employed to analyze, for example, the relative frequency of specific POS tags, as demonstrated in [EKHW23].
  Cited by 11 papers: [RL21, EKHW23, ARD24, AM24, DS24, JGM24, T$^+$23, GKSJ21, LCZ21, PLLGMP21, SASAW20]

- **Parse Tree height:** This feature measures the average height of parse trees, providing insights into sentence complexity.
  Cited by 1 paper: [EKHW23]

- **Personal Pronouns:** The frequency of personal pronouns such as "we," "he," and "I" can be analyzed, with potential subdivisions into first- and third-person pronouns, as explored in [EKHW23].
  Cited by 6 papers: [RK20a, EKHW23, JGM24, SPR22, ASM22, SGS21]

- **Gender and Group Identification:** This feature assesses whether the post references specific genders or groups, identifiable through the use of terms like "he" or "she" (e.g., in "She is sooooo pretty") using a POS tagger [LCZ21].
  Cited by 2 papers: [LCZ21, GKSJ21]

- **Function Words :** The occurrences of function words, including particles, prepositions, auxiliary verbs, and modal verbs, are encoded by this feature.
  Cited by 3 papers: [RK20a, SGS21, PLLGMP21]

- **Tense:**This feature captures the occurrence of grammatical structures associated with past, present, or future tenses.
  Cited by 10 papers: [SGS21]

- **Readability Metrics:** Metrics such as the Automated Readability Index (ARI) that assess the complexity of a text by examining factors like average characters per word and sentence length [RK20a]. Or the Flesch Reading Ease formula [Fle48], which evaluates average sentence length and syllables per word, is another example.
  Cited by 5 papers: [RK20a, EKHW23, JGM24, AHB$^+$23, CKB$^+$21]

- **Formality Metrics:** the degree of formality present in the text.
  Cited by 1 paper: [EKHW23]

- **Sentence / Word Length and density:** measured for example in the average or variance of sentence and word lengths, which can contribute to readability metrics or be treated as independent features. Another example would be measuring the word density (e.g., the number of words per 100 characters).
  Cited by 7 papers: [RK20a, HLM18, EKHW23, JGM24, AHB$^+$23, SGS21, SCM$^+$17]

Semantic Features

These features relate to the underlying meaning of a post, such as the base the emotion.

- **Sentiment / Polarity:** This feature captures positive, negative, or neutral emotions. Basic sentiments can be further divided into subcategories, such as anger, anxiety, and sadness, as shown in [ARD24]. Individual word sentiments can be obtained from dictionaries. Sentiment can be captured as categories or, in cases of ambiguity, as the ratio of positive to negative words [GKSJ21].
  Cited by 12 papers: [RK20a, HLM18, EKHW23, ARD24, KWR+24, AHB+23, SPR22, ASM22, GKSJ21, LCZ21, SCM+17, BW15]
- **Named Entity Recognition (NER):** can capute the mentions of entities such as organizations, locations, or people in a post, by applying, for example, a one-hot encoding of the entities [EKHW23]. It can also measure how many NER instances certain posts share, contributing to the construction of a similarity metric, as done in [HP20].
  Cited by 2 papers: [EKHW23, SASAW20]
- **Topic Modeling:** This technique identifies latent themes, commonly referred to as topics, using methods such as Latent Dirichlet Allocation (LDA).
  Cited by 3 papers: [ARKL18, ZSKP21, KA24]

Context Features

These features do not directly deal with the post itself but rather its context and the circumstances under which it created.

- **Publication Time:** captures when the content was posted.
  Cited by 3 papers: [ARKL18, HLM18, BW15]
- **Publication Time Rank:** AThis indicates the order in which the content was published (e.g., as the second post or the 105th).
  Cited by 1 paper: [HLM18]
- **Quote/Reply:** identifies whether the new post quotes previous content or is a response to another post.
  Cited by 1 paper: [HLM18]
- **Weather Information:** includes environmental factors at the time of posting, such as temperature and humidity.
  Cited by 1 paper: [ARKL18]
- **Competing content:** assesses, for example, the number of similar posts (under an article) available at the time of publication.
  Cited by 1 paper: [ARKL18]

- **Author Information:** encompasses various details about the user or publisher, such as follower count [MS23], post volume [SPR22], account creation date [CKB+21], and common topics of discussion [CKB+21].
  Cited by 6 papers: [ARKL18, MS23, SPR22, CKB+21, CKB+21, NPPZ18]

Other features

Embeddings are mentioned here as a special category due to their comparatively lower explainablity compared to other features. They represent high-dimensional contextual representations of words or phrases and are increasingly employed in text data modeling.

- **Word Embeddings:** transform words into dense vector representations based on their context. These embeddings can be pretrained or specifically trained for a task. Notable examples include Doc2Vec, Word2Vec (developed by Google), GloVe (from Stanford), and fastText.
  Cited by more than 10 papers: [ARKL18, WRJB18, ANM+21, RL21, KASA24, T+23, SD22, MW20]

Insights

A notable insight from the literature review is the divergent approaches to feature selection during preprocessing. In some studies, specific parts of a text are removed without any assessment of their significance, while in others, these same parts are seen as valuable features. For example, in [LCZ21], punctuation, numbers, and emoticons are discarded, whereas many other studies utilize these elements in model training. Another common practice is the removal of stop words; for instance, while some studies advocate for their exclusion, others highlight the importance of the ratio of filler words to relevant words, noting that fake news articles tend to contain fewer stop words than credible ones [SGS21]. This suggests a potential lesson: features such as filler words should not be discarded prematurely but rather evaluated and properly encoded as new features before determining their relevance.

CHAPTER 5

Experimental Setup: Popularity Prediction

The primary objective of this research is to explore the task of predicting the popularity of posts. This will be approached through both regression analysis [CH15], where the response variable is continuous, and classification tasks, where the output is a discrete variable.

- **Regression Tasks:**
  - *Total Number of Upvotes*
  - *Total Number of Downvotes*
  - *Relative Popularity:*

    The ratio of upvotes to downvotes, where 1 represents a post with only upvotes, 0 a post with only downvotes, and 0.5 a post with an equal number of upvotes and downvotes.
- **Classification Task:**
  - *Type of Comment:*

    This feature will have values "Top Post" or "Regular Post". The top $n\%$ of posts under a certain article, based on engagement (sum of upvotes and downvotes), will be labeled as "Top Posts". Initially, n will be set to 10%. After further analysis, the most suitable percentile will be determined and applied.

Features for Prediction

Determining the right data to feed into a statistical model to predict a certain outcome is a central task in natural language processing (NLP). Raw data from datasets is often insufficient for effective modeling without transformation. As a result, the literature discusses approaches to obtain the right input for prediction, such as *feature engineering*, which involves selecting the right subset of informative features or combining existing features to create new ones [GB12]. Or *feature generation* by utilizing domain-specific and common-sense knowledge [GM05].

This research will utilize two base sets of features to train a set of explainable Machine Learning (ML) models, such as Random Forests [SLT+03], Linear Regression, and Logistic Regression models, to predict post popularity and establish baseline solutions. These baselines are designed to incrementally build upon each other, offering a structured approach to exploring feature sets and increasing model complexity.

- **Baseline 1: Information on Previous Votes**
  This baseline model utilizes only the median value of the *PositiveVotes* and *NegativeVotes* (of the posts in the training set) for prediction. The baseline, therefore, does not incorporate any text-related attributes or linguistic characteristics, serving as the absolute minimal benchmark for prediction performance.

- **Baseline 2: Meta-Text Features**
  Building upon Baseline 1, this model introduces features related to the text's structure and metadata without considering the actual content. Potential features include *BodyWordLength*, *HeadlineWordLength*, *BodyLength*, *HeadlineLength*, and the time elapsed since the text's creation. This baseline aims to capture meta-data about the text (like its volume or when it was created), providing a richer set of features while still avoiding content-specific information.

- **Baseline 3: Simple Content Features - Rule-based Solution**
  This baseline incorporates simple linguistic features derived from the text using rule-based methods. Examples of such features include the number of spelling mistakes, the presence of conjunctive words, occurrences of questions, and expressions of gratitude. This model aims to capture more nuanced linguistic aspects that may influence the reception of the post. The specific rule-based features to be used will be subject to experimentation.

These baseline models will be evaluated to establish performance benchmarks. The layering of these baselines allows for a systematic exploration of feature sets, starting from basic numerical data to more complex linguistic features. After assessing these foundational approaches, more advanced NLP methods will be explored.

Model Training and Evaluation

The dataset will be randomly split into training and test sets, in a ratio of 80 to 20% (a common ratio of a train/test split in this research field). The training set will be used to train the models, while the test set will be used to evaluate their performance.

Quantitative Analysis

For the regression tasks, model performance will be assessed using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For the classification task, precision, recall and F1 score will be used as evaluation metrics.

Qualitative Analysis

In addition to quantitative metrics, qualitative analysis will be conducted to interpret the findings. This involves analyzing model predictions to understand why certain posts gain popularity, become unpopular, or evoke polarization. The insights gained will be contextualized within the broader literature.

CHAPTER 6

Results

CHAPTER 7

Discussion

CHAPTER 8

Conclusion, Limitations & Future Work

List of Figures

List of Tables

List of Algorithms

Bibliography

[AA19] Hissah ALSaif and Taghreed Alotaibi. Arabic text classification using feature-reduction techniques for detecting violence on social media. *International Journal of Advanced Computer Science and Applications*, 10(4), 2019.

[AGAR+22] Abdullah Marish Ali, Fuad A Ghaleb, Bander Ali Saleh Al-Rimy, Fawaz Jaber Alsolami, and Asif Irshad Khan. Deep ensemble fake news detection model using sequential deep learning technique. *Sensors*, 22(18):6970, 2022.

[AHB+23] Anuja Arora, Vikas Hassija, Shivam Bansal, Siddharth Yadav, Vinay Chamola, and Amir Hussain. A novel multimodal online news popularity prediction model based on ensemble learning. *Expert Systems*, 40(8):e13336, 2023.

[AM24] Syed Fahad Ali and Nayyer Masood. Evaluation of adjective and adverb types for effective twitter sentiment classification. *Plos one*, 19(5):e0302423, 2024.

[ANM+21] Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. Rp-mod rp-crowd: Moderator-and crowd-annotated german news comment datasets. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

[ARD24] Keshopan Arunthavachelvan, Shaina Raza, and Chen Ding. A deep neural network approach for fake news detection using linguistic and psychological features. *User Modeling and User-Adapted Interaction*, 34(4):1043–1070, 2024.

[ARKL18] Carl Ambroselli, Julian Risch, Ralf Krestel, and Andreas Loos. Prediction for the newsroom: Which articles will get the most comments? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 193–199, 2018.

[ASM22] Fatimah Alkomah, Sanaz Salati, and Xiaogang Ma. A new hate speech detection system based on textual and psychological features. *Int J Adv Comput Sci Appl.*, 13(8):860–869, 2022.

[BAH12] Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 26–33, 2012.

[BM12] Pablo J Boczkowski and Eugenia Mitchelstein. How users take advantage of different forms of interactivity on online news sites: Clicking, e-mailing, and commenting. *Human communication research*, 38(1):1–22, 2012.

[BW15] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242, 2015.

[CH15] Samprit Chatterjee and Ali S Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

[CKB+21] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, Mario Navarro, et al. Predicting age groups of reddit users based on posting behavior and metadata: classification model development and validation. *JMIR Public Health and Surveillance*, 7(3):e25807, 2021.

[CPKG16] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 9–16. IEEE, 2016.

[DGVM20] Ishita Daga, Anchal Gupta, Raj Vardhan, and Partha Mukherjee. Prediction of likes and retweets using text information retrieval. *Procedia computer science*, 168:123–128, 2020.

[DS24] Shilpa Dixit and Nitasha Soni. Enhancing stock market prediction using three-phase classifier and em-epo optimization with news feeds and historical data. *Multimedia Tools and Applications*, 83(13):37859–37887, 2024.

[DWW19] Keyan Ding, Ronggang Wang, and Shiqi Wang. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2682–2686, 2019.

[EKHW23] Elisabeth Eder, Ulrike Krieg-Holz, and Michael Wiegand. A question of style: A dataset for analyzing formality on different levels. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 580–593, 2023.

[Fle48] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[GB12] Vijay N Garla and Cynthia Brandt. Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics*, 45(5):992–998, 2012.

[GIMW11] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 581–586, 2011.

[GKSJ21] R Geetha, S Karthika, Chaluvadi Jwala Sowmika, and Bharathi M Janani. Auto-off id: Automatic detection of offensive language in social media. In *Journal of Physics: Conference Series*, volume 1911, page 012012. IOP Publishing, 2021.

[GM05] Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, volume 5, pages 1048–1053, 2005.

[GS23] Şura Genç and Elif Surer. Clickbaittr: Dataset for clickbait detection from turkish news sites and social media with a comparative analysis via machine learning algorithms. *Journal of Information Science*, 49(2):480–499, 2023.

[HLM18] Marlo Häring, Wiebke Loosen, and Walid Maalej. Who is addressed in this comment? automatically classifying meta-comments in news comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–20, 2018.

[HP20] Jacek Haneczok and Jakub Piskorski. Shallow and deep learning for event relatedness classification. *Information Processing & Management*, 57(6):102371, 2020.

[JGM24] Mayank Kumar Jain, Dinesh Gopalani, and Yogesh Kumar Meena. Confake: fake news identification using content based features. *Multimedia Tools and Applications*, 83(3):8729–8755, 2024.

[KA24] M Kavitha and K Akila. Amplifying document categorization with advanced features and deep learning. *Multimedia Tools and Applications*, pages 1–19, 2024.

[KASA24] Muhammad Kamran, Ahmad S Alghamdi, Ammar Saeed, and Faisal S Alsubaei. Mr-fnc: A fake news classification model to mitigate racism. *International Journal of Advanced Computer Science & Applications*, 15(2), 2024.

[Kit04] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[KWR+24] Akib Mohi Ud Din Khanday, Mudasir Ahmad Wani, Syed Tanzeel Rabani, Qamar Rayees Khan, and Ahmed A Abd El-Latif. Hapi: An efficient hybrid feature engineering-based approach for propaganda identification in social media. *Plos one*, 19(7):e0302583, 2024.

[LCZ21] Cheng-Te Li, Hsin-Yu Chen, and Yang Zhang. On exploring feature representation learning of items to forecast their rise and fall in social media. *Journal of Intelligent Information Systems*, 56(3):409–433, 2021.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LWX+22] An-An Liu, Xiaowen Wang, Ning Xu, Junbo Guo, Guoqing Jin, Quan Zhang, Yejun Tang, and Shenyuan Zhang. A review of feature fusion-based media popularity prediction methods. *Visual Informatics*, 6(4):78–89, 2022.

[LZZ20] Xin Lai, Yihong Zhang, and Wei Zhang. Hyfea: winning solution to social media popularity prediction for multimedia grand challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4565–4569, 2020.

[MCCH22] Yi-Wei Ma, Jiann-Liang Chen, Li-Dong Chen, and Yueh-Min Huang. Intelligent clickbait news detection system based on artificial intelligence and feature engineering. *IEEE Transactions on Engineering Management*, 2022.

[MKR+24] Muhammad Mujahid, EROL Kına, Furqan Rustam, Monica Gracia Villar, Eduardo Silva Alvarado, Isabel De La Torre Diez, and Imran Ashraf. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, 11(1):87, 2024.

[MS23] Sahar Mehravaran and Pirooz Shamsinejadbabaki. Devising a machine learning-based instagram fake news detection system using content and context features. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 47(4):1657–1666, 2023.

[MW20] Zewdie Mossie and Jenq-Haur Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087, 2020.

[NKS21] Maria N Nelson, Thomas B Ksiazek, and Nina Springer. Killing the comments: Why do news organizations remove user commentary functions? *Journalism and Media*, 2(4):572–583, 2021.

[NPPZ18] Paolo Nesi, Gianni Pantaleo, Irene Paoli, and Imad Zaza. Assessing the retweet proneness of tweets: predictive models for retweeting. *Multimedia tools and applications*, 77(20):26371–26396, 2018.

[PLLGMP21] Gabriel Ichcanziho Pérez-Landa, Octavio Loyola-González, and Miguel Angel Medina-Pérez. An explainable artificial intelligence model for detecting xenophobic tweets. *Applied Sciences*, 11(22):10801, 2021.

[PSDE16] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1114–1125, 2016.

[RK20a] Julian Risch and Ralf Krestel. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 579–589, 2020.

[RK20b] Julian Risch and Ralf Krestel. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, pages 85–109, 2020.

[RL21] Surangika Ranathunga and Isuru Udara Liyanage. Sentiment analysis of sinhala news comments. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–23, 2021.

[SASAW20] Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, and Bianca Wright. Sarcasm detection using machine learning algorithms in twitter: A systematic review. *International Journal of Market Research*, 62(5):578–598, 2020.

[SCM+17] Abeed Sarker, Pramod Chandrashekar, Arjun Magge, Haitao Cai, Ari Klein, and Graciela Gonzalez. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *Journal of medical Internet research*, 19(10):e361, 2017.

[SD22] R Sandrilla and M Savitha Devi. Fnu-bicnn: Fake news and fake url detection using bi-cnn. *International Journal of Advanced Computer Science and Applications*, 13(2), 2022.

[SGS21] Vivek K Singh, Isha Ghosh, and Darshan Sonagara. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17, 2021.

[SLT+03] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[SPR22] Maya Stemmer, Yisrael Parmet, and Gilad Ravid. Identifying patients with inflammatory bowel disease on twitter and learning from their personal experience: retrospective cohort study. *Journal of medical Internet research*, 24(8):e29186, 2022.

[SST17] Dietmar Schabus, Marcin Skowron, and Martin Trapp. One million posts: A data set of german online discussions. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1241–1244, 2017.

[STT+20] Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*, 2020.

[T+23] P Santhi Thilagam et al. Multi-layer perceptron based fake news classification using knowledge base triples. *Applied Intelligence*, 53(6):6276–6287, 2023.

[TWDR09] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1765–1768, 2009.

[WRJB18] Gregor Wiedemann, Eugen Ruppert, Raghav Jindal, and Chris Biemann. Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. *arXiv preprint arXiv:1811.02906*, 2018.

[ZSKP21] Elaine Zosa, Ravi Shekhar, Mladen Karan, and Matthew Purver. Not all comments are equal: Insights into comment moderation from a topic-aware model. *arXiv preprint arXiv:2109.10033*, 2021.