Public Reaction to Scientific Research via Twitter Sentiment Prediction

Murtuza Shahzad and Hamed Alhoori

Department of Computer Science, Northern Illinois University, DeKalb, IL, USA

Abstract

Purpose: Social media users share their ideas, thoughts, and emotions with other users. However, it is not clear how online users would respond to new research outcomes. This study aims to predict the nature of the emotions expressed by Twitter users toward scientific publications. Additionally, we investigate what features of the research articles help in such prediction. Identifying the sentiments of research articles on social media will help scientists gauge a new societal impact of their research articles.

Design/methodology/approach: Several tools are used for sentiment analysis, so we applied five sentiment analysis tools to check which are suitable for capturing a tweet's sentiment value and decided to use NLTK VADER and TextBlob. We segregated the sentiment value into negative, positive, and neutral. We measured the mean and median of tweets' sentiment value for research articles with more than one tweet. We next built machine learning models to predict the sentiments of tweets related to scientific publications and investigated the essential features that controlled the prediction models.

Findings: We found that the most important feature in all the models was the

Email addresses: msyed1@niu.edu (Murtuza Shahzad), alhoori@niu.edu (Hamed Alhoori)

sentiment of the research article title followed by the author count. We observed that the tree-based models performed better than other classification models, with Random Forest achieving 89% accuracy for binary classification and 73% accuracy for three-label classification.

Research Limitations: In this research, we used state-of-the-art sentiment analysis libraries. However, these libraries might vary at times in their sentiment prediction behavior. Tweet sentiment may be influenced by a multitude of circumstances and is not always immediately tied to the paper's details. In the future, we intend to broaden the scope of our research by employing word2vec models.

Practical Implications: Many studies have focused on understanding the impact of science on scientists or how science communicators can improve their outcomes. Research in this area has relied on fewer and more limited measures, such as citations and user studies with small datasets. There is currently a critical need to find novel methods to quantify and evaluate the broader impact of research. This study will help scientists better comprehend the emotional impact of their work. Additionally, the value of understanding the public's interest and reactions helps science communicators identify effective ways to engage with the public and build positive connections between scientific communities and the public.

Originality/value: This study will extend work on public engagement with science, sociology of science, and computational social science. It will enable researchers to identify areas in which there is a gap between public and expert understanding and provide strategies by which this gap can be bridged.

Keywords: Sentiment Analysis, Social Media, Twitter, Emotional Impact, Public Understanding of Science, Science and Technology Studies

---

1. Introduction

Microblogging has become ubiquitous, and the scope of the texts posted on microblogging platforms and their efficacy as a means of communication has far exceeded expectations. Social media platforms have become a place where users collaborate, share their ideas and also have conflicts (Hansson et al., 2019; Hansson and Ludwig, 2019). With 126 million active daily users (Shaban, 2019), Twitter is the dominant microblogging platform on which users discuss a breadth of subjects and even play a role in influencing current trends. Users on Twitter post short and often informal messages (tweets) in which they share information and project opinions and sentiments about what is going on in the world. Twitter has been a major platform for sharing scholarly articles, and many researchers have used it to develop various metrics for scholarly articles (Haustein, 2019). Other social media platforms like Facebook and Weibo have also been sources to study online users' responses (Kou et al., 2017). Social media platforms have become a hub where users express their opinions and emotions related to multiple fields of interest (Chatterjee et al., 2019). Researchers have studied the sentiments and emotions associated with research articles on these platforms (Freeman et al., 2019, 2020).

There is a need to understand the impact of research beyond the traditional scholarly impact (i.e., citations)(Le et al., 2019; Kousha and Thelwall, 2019; Noyons, 2019) such as the impact on economics (Shaikh and Alhoori, 2019), public policy (Kale et al., 2017), social media (Alhoori et al., 2019), news outlets (Siravuri and Alhoori, 2017), and public understanding of science (Siravuri et al., 2018). In the present study, we subjected a collection of tweets to the process of sentiment analysis, which refers to the contextual mining of texts through which subjective information is identified

and extracted (Liu, 2012). Such subjective information is essential in many business-related opinion-mining contexts. For scholarly literature, tweets can be analyzed to determine the popularity of a research article, whether it is liked, or even whether anyone has shared and discussed the article online. We analyzed a collection of tweets to identify whether tweets about a given research article were predominantly positive, neutral, or negative. We built machine learning models to predict the nature of the sentiments expressed in tweets about a given research article. We considered several distinct evaluative measures for the machine learning models and found that the tree-based models performed better than the other models in predicting tweet sentiments.

Knowing how social media data can be utilized to learn about a research article's emotional impact is an interesting quest. Such a study can pave the way for scientists to understand the impact their work could have based on the reactions to previous studies. In addition, they would know about the specific features in their paper that could be modified to avoid a negative reaction. To understand and analyze this impact, we consider the following research questions:

1. **RQ1.** Can we build machine learning models to predict the tweet sentiment for research articles?
2. **RQ2.** What are the crucial features that help in accurately predicting the sentiment of the tweet?

In summary, our contributions include

1. Analyzing tweets related to research articles using various social media features and research domains.

2. Understanding the sentiment of tweets for research articles using various state-of-the-art sentiment analysis libraries.
3. Building machine learning models to predict sentiments of tweets related to scientific publications.

2. Related Work

Twitter is one of the social media platforms on which many sentiment analysis and predictive models are built (Jaidka et al., 2021; Haunschild et al., 2020; Ibrahim and Wang, 2019; Didegah et al., 2018). In numerous studies, researchers have endeavored to predict the sentiment of general tweets and tweets related to specific areas or events. Some researchers have analyzed the sentiments of tweets related to scholarly articles and have built models to predict sentiments of tweets for scholarly articles for specific research domains (Hassan et al., 2020). Bharathwaj et al. (2019) predict the positive, negative, and neutral sentiment of tweets used for research articles in Medicine and Psychiatry Disciplines. They manually labeled 1,099 negatives, 2,000 positives, and 8,000 neutral tweets, built several machine learning models, and identified the best model with the Support Vector Machines (SVM) having 91.6% accuracy.

To estimate the polarity of tweets, Narr et al. (2012) subjected Twitter data to a language-independent sentiment analysis. They collected tweets in several languages and used the Naive Bayes classifier on the n-gram features to classify the sentiments. The mixed four-language unigram had an accuracy of 71.5%. Similarly, Bae and Lee (2012) studied the polarity and sentiments of tweets posted by celebrities with more than a million followers. The researchers performed a lexical sentiment analysis based on which the sentiment score was calculated. According to the different

types of correlational analysis, when a celebrity posted a positive tweet, his/her followers did likewise. Hence, based on tweet similarity, it appears that the celebrities influenced their respective followers. Kharde et al. (2016) tested various classification methods such as SVM, Naive Bayes, and Maximum Entropy using a Twitter dataset. The baseline model was the least accurate of all the classification models. Compared with that model, Maximum Entropy and Naive Bayes each returned a higher accuracy. However, the highest accuracy was achieved using SVM with Unigram and Bigram with stop word removal. In addition to these classification methods, the classification algorithm Naive Bayes was run separately as a Unigram Multinomial Bayes and a Multigram Multinomial Bayes (Parikh and Movassate, 2009). The results showed that between Maximum Entropy and Naive Bayes classification models, based on the unique nature of the tweets, Maximum Entropy could not take advantage of the sequenceable features. Hence each of the Naive Bayes classifiers performed overwhelmingly better than the other classification models did.

There were different approaches and methods applied in various other studies on Twitter. Zaman et al. (2010) suggested a probabilistic collaborative filter model that predicts future retweets, thereby showing the spread of information. Hao et al. (2011) also used a visual-based approach compared to the previous text-based approaches. The approaches discussed thus far rely on traditional methods to pre-process data. However, Da Silva et al. (2014) suggested that data can be preprocessed instead, using feature hashing in relation to a bag of words and lexicons, which are then fed into the classification algorithms. Comparing this method with the baseline method of bootstrapping an ensemble framework as used by Hassan et al. (2013), the researchers obtained similar accuracy, suggesting that classifier ensembles can be useful

in tweet sentiment analysis. On a similar note, Saif et al. (2014) used several methods—Zipf's law, term-based random sampling, mutual information, and the classic (pre-compiled) approaches—in conjunction with each other to remove the stop-words during preprocessing. In comparison with the baseline model results in which the stop words were not removed, their method reduced the feature space by nearly 65%, thereby maintaining high performance on classification by decreasing data sparsity up to 0.37%.

Likewise, Pak and Paroubek (2010) performed a linguistic analysis of tweets collected for their research. After preprocessing the tweets, the researchers extracted features that were then used in the Multinomial Naive Bayes classifier algorithm to predict positive, neutral, and negative sentiments. According to their results, bigrams are better than both unigrams and trigrams in capturing sentiment expression. Instead of considering all the text comprising a tweet, Kouloumpis et al. (2011) suggested analyzing only the hashtags present in the tweets to create what they referred to as a hashtagged dataset. On comparing the average accuracy of various features in predicting tweet sentiment, they found hashtags and emoticons to be more useful than part-of-speech features. Wang et al. (2011) developed that research direction further by analyzing only the hashtags of tweets and also boosting the graph-based classification algorithms, Loopy Belief Propagation (LBP), Relaxation Labeling (RL), and the Iterative Classification Algorithm (ICA). The researchers compared their model with the baseline model, which relied on traditional classification algorithms. Their results show that boosting achieved better results for predicting positive and negative tweets based on hashtags in terms of accuracy, precision, and recall compared with plain classification algorithms.

Wang et al. (2012b) used Semantic Role Labeling (SRL) to extract event-based tweets. The researchers used Latent Dirichlet Allocation (LDA) to identify the salient topics in the events. Using these topics, they built a model to predict future criminal events. In a similar attempt to predict crimes, Chen et al. (2015) used weather as a feature in addition to Twitter sentiment to predict the times and locations of crimes. Based on Kernel Density Estimation (KDE) used in conjunction with lexicon-based methods, they observed that temperature, aggression, and crime rate are related: high temperatures were associated with a high level of aggression, which ultimately led to a higher crime rate than when temperatures were low. Furthermore, in the context of weather, Mandel et al. (2012) considered tweets selected based on the level of concern expressed and demographic information related to Hurricane Irene. The researchers found that the number of tweets related to the hurricane directly affected the region peaks at the time of the hurricane and that the level of concern expressed in the tweets depended on the particular region.

In addition to crime and weather, the literature includes analyses of many other topics ranging from the incidence of disease to the stock market. For example, Achrekar et al. (2011) analyzed the content of tweets in an endeavor to predict flu trends. In relation to the stock market, Mittal and Goel (2012) drew on Twitter sentiments to analyze stock market trends. The researchers subjected the tweets to natural language processing and then to several classification models. The researchers found that Self Organizing Fuzzy Neural Networks returned the most accurate results.

The literature includes several studies indicating that tweets can influence election results. Pal et al. (2018) show how politicians may benefit from antagonistic messaging. Wang et al. (2012a) used a real-time data-processing infrastructure on IBM's

InfoSphere stream platform to write visualization modules and perform an analysis of the tweets. In a special case study, Bermingham and Smeaton (2011) analyzed the relationship between emotions expressed in tweets and election results for the Irish general election of 2011. Of the studies published to date, this research is unique in that it differentiates between the emotions expressed in tweets pertaining to the Inter-Party polls and the emotions expressed in tweets pertaining to the Intra-Party polls by implementing a separate measure for each. According to the results, during the weeks before the election, the sentiments were close to each other, but the sentiments were polarized on the day before the election. A model developed by Gayo-Avello (2012) predicted that not only would Barack Obama win the US Presidential Election of 2008 but that he would also win all the states. Overall, given that Obama both won the election and all the states with the exception of Texas, the results of the model were very accurate. However, the Texas election result also indicated that such models should be used with caution, given that the tweet sentiments cannot entirely predict the outcome of an election.

According to several research studies, when used with sentiment analysis and the addition of natural language processing, machine learning models can generate relatively accurate results. Exploring further, Neethu and Rajasree (2013) used SVM to train the machine to calculate and predict the sentiment scores of an electronic product. Focusing on movie reviews, Amolik et al. (2016) used machine learning and sentiment scores via a unigram approach to preprocessing the data. The researchers also ensured that hashtags were removed and then stored in the feature vectors. They used both SVM and Naive Bayes to predict the sentiments of movie reviews and then compared the scores to the baseline model on the metrics of precision and recall.

In this paper, we built predictive models to determine the sentiments expressed about research articles in tweets. There are a number of differences between our study and previous work. For example, we used a large dataset in our study compared with other studies. We use tweets related to research articles belonging to a broader range of scientific domains to observe what domains are better indicators of sentiment prediction. Additionally, we did not manually label the sentiment of tweets. Instead, we analyzed different sentiment analysis libraries to obtain class labels that were used as dependent variables in our machine learning models. For class imbalance issues among positive, negative, and neutral sentiments, we used state of the art class imbalance technique called Synthetic Minority Oversampling Technique (SMOTE). Further, some studies have either eliminated or included neutral sentiments for predictions. We built models with and without neutral sentiments to observe how the predictions work.

3. Data

We used Altmetrics data released from Altmetric.com in July 2018. Altmetrics (Akella et al., 2021; Alhoori and Furuta, 2014; Alhoori et al., 2014, 2015), consist of mentions of scholarly articles in online social media, such as Facebook, Twitter, and Wikipedia, and in online reference managers, such as Mendeley. The 2018 release of the Altmetric dataset consists of the details about the online mentions of about 19 million publications. The data comprise details about research articles, including a given article's title, author(s), and subject(s), as well as tweets about the article and the number of times it has been shared on social media. In this study, we focused on the research articles shared on Twitter to determine the tweets' sentiments. To meet

this goal, we first filtered the entire Altmetric dataset having the details of tweets. With this filtering, we were left with 6,011,003 articles. We took a random sample of 150,000 research articles from the 6 million articles using the Pandas library (McKinney et al., 2011). This random sampling was done by selecting a certain number of articles (150,000) without replacement from the dataset. Finally, we eliminated all records with missing values. The final dataset had altmetrics for 148,712 research articles with a total of 1,941,348 tweets. The entire study was done on these randomly selected articles. To the best of our knowledge, this is one of the largest datasets in such studies. The altmetrics features used in this study for data analysis and for predicting the tweets' sentiments are mentioned in Table 1. Some of the features we used, such as the abstract length of the article and the number of followers of the Twitter user, were shown to be important factors in measuring the popularity of articles on Twitter (Pandian et al., 2019).

Table 1: Selected features from the Altmetrics dataset.

Feature	**Description**
Scopus subject	Subject of a research article.
Article title	Title of a research article.
Article abstract	Abstract of a research article.
Abstract length	Number of words in the abstract of a research paper.
Follower count	Number of followers a Twitter user has.
Author count	Number of authors credited on the research article.
Tweet	Tweet about a research article.

From the features listed in Table 1, we derived some new features in Table 2. The final features used for the machine learning models are title sentiment, abstract sentiment, abstract length, tweet reach, author count, and tweet sentiment. Given that tweet sentiment is the target variable, we developed machine learning models to

Table 2: Derived features from the dataset.

Original feature	**Derived feature**	**Description**
Article title	Title sentiment	Sentiment score of the title of a research article.
Article abstract	Abstract sentiment	Sentiment score of a research article abstract.
Follower count	Tweet reach	The mean number of followers of each user who tweeted about the research article (i.e., one article can be tweeted by many users, who may differ from each other in the number of followers they have).
Tweet	Tweet sentiment	Sentiment score of a tweet related to a research article.

predict this variable.

We used open-source libraries such as Pandas and Matplotlib to load, manipulate, analyze, and visualize the data. We plotted graphs showing various insights. We first observed the number of online mentions of research articles on Twitter from 2011 to 2017. Figure 1 shows an increase in the number of articles shared on Twitter in this timespan. As the dataset did not have the altmetrics for the entire year of 2018, we removed this year from the plot to avoid misinterpretation because of the decrease in tweets for 2018. The number of tweets in the year 2018 (until July) was 293,881.

We then extracted the Scopus subjects from the dataset to determine the extent of their popularity on Twitter. Figure 2 shows the number of tweets for each subject, and Figure 3 shows the number of articles for each subject. We observed that Health Sciences had the most articles and most tweets, followed by Medicine.

We obtained 30 Scopus subjects[1] from the dataset, which have four higher-level groupings and 27 lower-level groupings of subjects. The four higher-level subjects are Physical Sciences, Health Sciences, Social Sciences, and Life Sciences. It is to be

[1] `https://service.elsevier.com/app/answers/detail/a_id/12007/supporthub/scopus/`

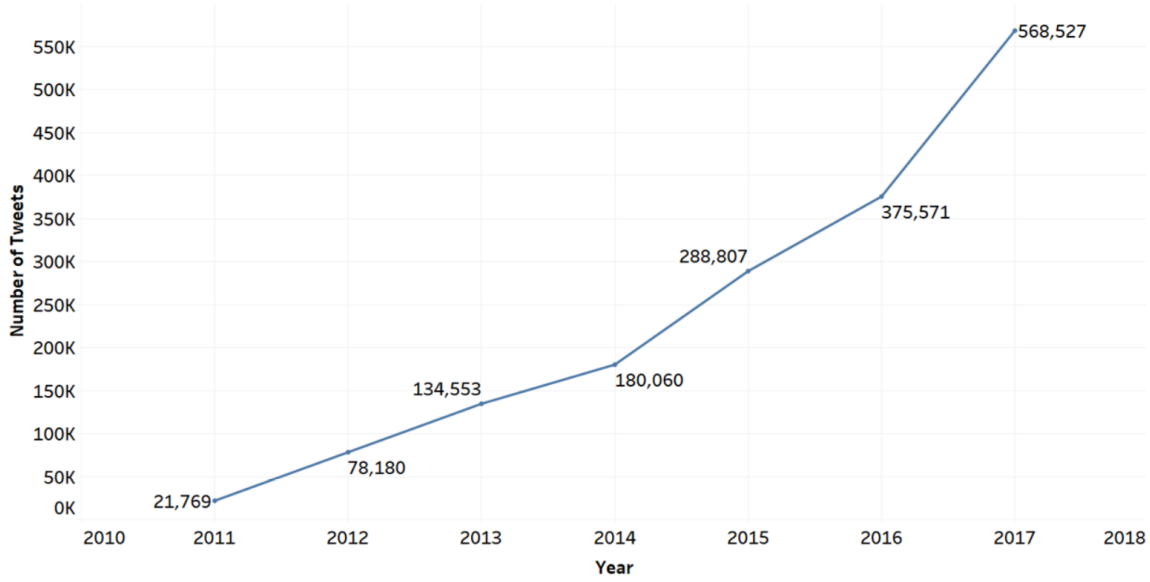


Figure 1: Number of tweets related to research articles for the years 2011-2017.

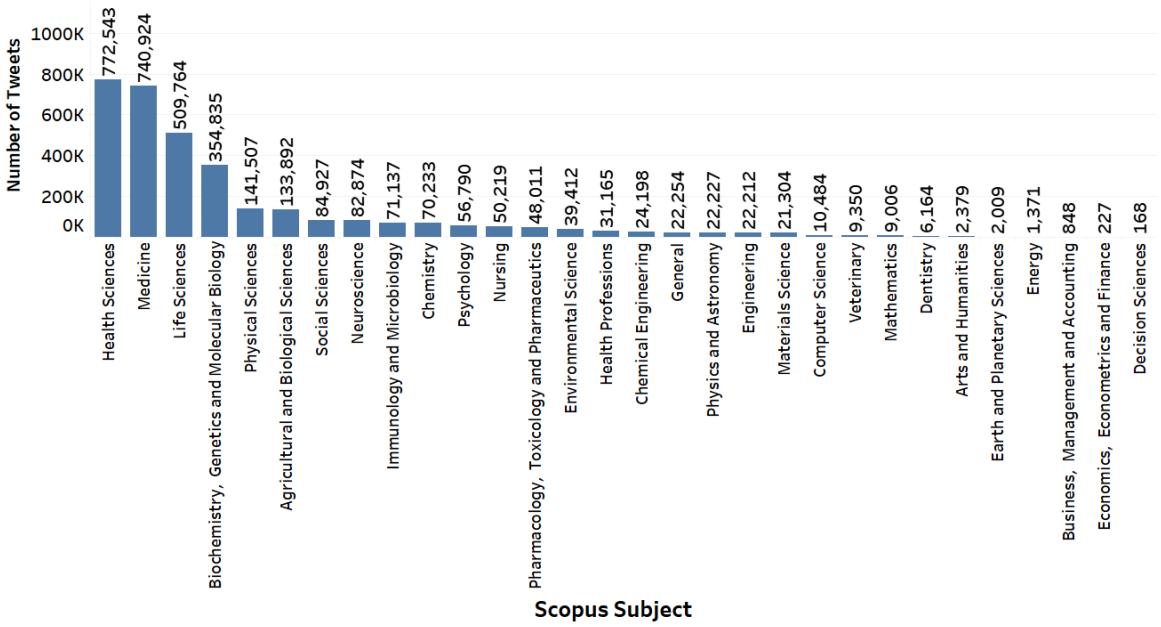


Figure 2: Number of tweets for each Scopus subject.

noted that the subject Social Sciences is in the higher and the lower-level groupings.

We noticed that the dataset from Altmetric categorizes Scopus subjects by merg-

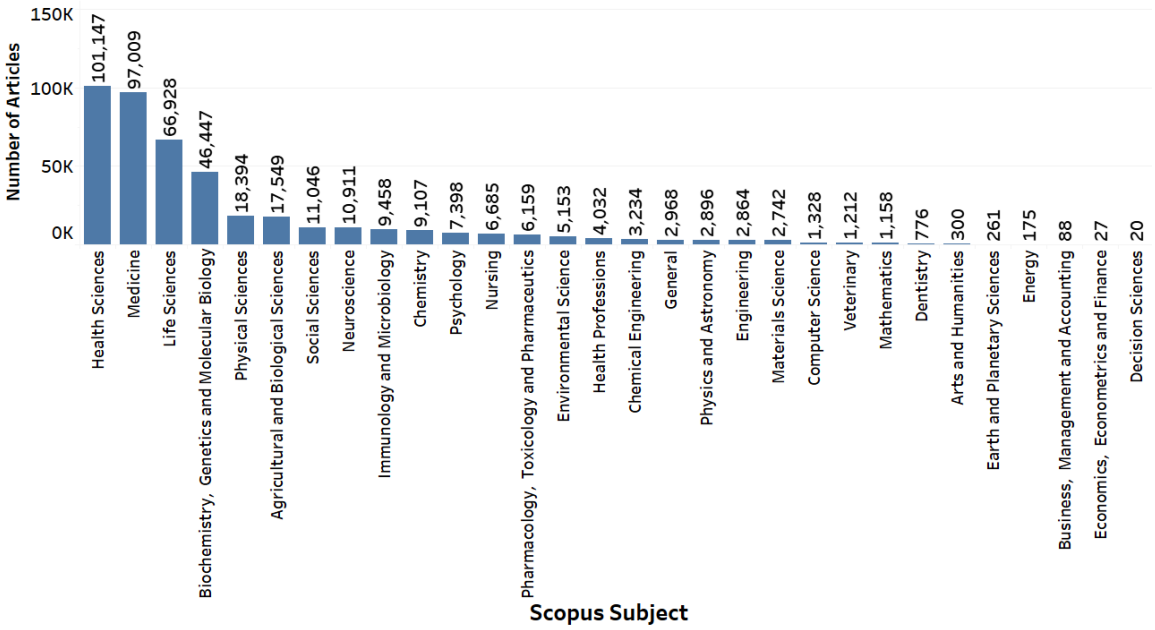


Figure 3: Number of articles for each Scopus subject.

ing the higher and lower level groupings of Scopus subjects. As an article may belong to multiple Scopus subjects, the Altmetric dataset also categorizes a single article to multiple Scopus subjects. For example, an article in the Altmetric dataset titled "Respiratory Factors Contributing to Exercise Intolerance in Breast Cancer Survivors: A Case-Control Study" falls under the category of Medicine, Health Sciences, and Nursing Scopus subjects.

Detecting the sentiment of any given text involves some challenges. Mohammad (2017) pointed out that the determination of sentiment could be at different text granularities such as sentiment of a word, a sentence, a paragraph, or an entire document. Mohammad discussed that another challenge is setting a threshold for negative, neutral, and positive sentiments and suggested that some applications may just require the detection of extremely positive and negative instances. Hussein (2018) found

that domain-dependence is another important component in recognizing sentiment. Figures of speech, semantics, explicit and implicit opinions, regular and comparative opinions in tweets are challenges for identifying sentiment analysis of tweets (Pozzi et al., 2017). In analyzing the sentiments in academic writing, Vinkers et al. (2015) found that the use of positive words like novel, robust, innovative, and unprecedented in the titles and abstracts of research articles published between 1974 and 2014 grew significantly. Negative words have also increased in frequency, albeit in a smaller but statistically significant way. For the research articles selected in our study, Table 3 shows the top 25 positive and negative words in the title, abstract, and tweets related to the articles given by the TextBlob sentiment analysis library.

4. Methods

To obtain sentiment scores for the tweets, the title of the research articles, and the abstracts of the research articles, we used the following Python libraries:

1. **NLTK**[2] **VADER** (Valence Aware Dictionary and sEntiment Reasoner) is a lexical and rule-based sentiment analysis tool in the Natural Language Toolkit library (NLTK). A component of VADER, the sentiment intensity analyzer generates a compound polarity score for text. This score is a continuous value in the range of -1 (negative) to +1 (positive). We used version 3.6.2 of the NLTK library.
2. **TextBlob**[3] is a library used for various purposes, such as part-of-speech tagging, noun phrase extraction, and sentiment analysis. This library also generates

[2] `https://www.nltk.org/`
[3] `https://pypi.org/project/textblob/`

Table 3: Top 25 positive and negative words in title, abstract, and tweets of research articles.

Title		Abstract		Tweets	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
best	boring	awesome	awful	awesome	awful
delicious	devastating	best	bleak	best	bleak
excellent	disgusting	delicious	boring	breathtaking	boring
greatest	evil	excellent	cruel	delicious	cruel
perfect	grim	exquisite	devastating	delightful	devastating
superb	vicious	flawless	disgusted	excellent	disgusting
wonderful	worst	greatest	dreadful	exquisite	dreadful
brilliant	fearful	impressed	evil	greatest	evil
ideal	repellent	legendary	grim	impressed	grim
incredible	retard	magnificent	gruesome	legendary	gruesome
beautiful	base	marvelous	horrible	magnificent	horrible
splendid	bloody	masterful	horrific	marvelous	horrific
attractive	doubtful	perfect	hysterical	masterful	hysterical
experienced	filthy	superb	insane	perfect	insane
expressive	grief	wonderful	insulting	priceless	insulting
favored	hate	artesian	menacing	superb	miserable
great	violent	brilliant	outrageous	wonderful	nasty
happy	stupid	ideal	ruthless	brilliant	outrageous
intelligent	tragic	incredible	shocking	ideal	pathetic
joy	sick	beautiful	terrible	incredible	shocking
proud	anger	attractive	terrifying	beautiful	terrible
uncommon	crude	brave	vicious	splendid	terrifying
unforgettable	frustrated	elect	worst	attractive	vicious
win	painful	experienced	fearful	brave	worst
remarkable	shocked	expressive	hated	elect	fearful

a sentiment score as a continuous value in the range of -1 (negative) to +1 (positive). We used version 0.15.3 of the TextBlob library.

3. **Stanford CoreNLP**[4] is a library built in Java. For Stanford CoreNLP, the Python packages interact with the library on a server running in the background

[4] `https://stanfordnlp.github.io/CoreNLP/`

on the Java platform. This library generates sentiment scores as follows: very negative = 0, negative = 1, neutral = 2, positive = 3, and very positive = 4. We used version 4.2.0 of the Stanford CoreNLP library.

4. **SentiStrength**[5] is an algorithm by Thelwall et al. (2010) used to extract sentiments of informal texts. We used a Python wrapper of this algorithm to generate ternary sentiments: negative = -1, neutral = 0, positive = 1.
5. **Sentiment140**[6] is a sentiment analysis tool specifically designed for Twitter. We used the sentiment140 API to query the tweets' sentiment and obtain the polarities as negative = 0, neutral = 2, and positive = 4.

Each article might receive multiple tweets (e.g., tweets by different users), and the sentiment of those tweets determine the positive, negative, or neutral impact of the research article. To get an overall sentiment value for a research article, we used the mean and median of multiple tweet sentiments for a particular research article. We looked at the sentiment scores these libraries gave to the tweets. If they gave a high level of neutral sentiments, we did not use those libraries in building machine learning models. The Stanford CoreNLP library generated a score of neutral for most tweets such that we could not use the scores in our study. Table 4 shows that SentiStrength and Sentiment140 libraries resulted in many neutral sentiments. Surprisingly, we observe that the SentiStrength library gave more negative sentiments than positive sentiments. Some studies (Friedrich et al., 2015a,b) have shown that applying the SentiStrength library to the tweets of scientific articles resulted in a high number of neutral sentiments of the tweets. However, for the purpose of this study,

[5] `http://sentistrength.wlv.ac.uk/`
[6] `http://www.sentiment140.com/`

it was appropriate to select those libraries that would generate more non-neutral sentiments. This would help us to build machine learning models that have less class imbalance. Therefore, as the Stanford CoreNLP, SentiStrength, and Sentiment140 libraries generated most of the tweets' sentiment as neutral, we did not use these libraries any further in our study.

Table 4: Sentiment distribution of articles using SentiStrength and Sentiment140 libraries.

Sentiment library	Metric for multiple sentiments	Number of positive sentiments	Number of negative sentiments	Number of neutral sentiments
SentiStrength	mean	11,443 ($\approx$ 7.7%)	31,212 ($\approx$ 21%)	106,057 ($\approx$ 71.3%)
SentiStrength	median	14,905 ($\approx$ 10%)	39,091 ($\approx$ 26.3%)	94,716 ($\approx$ 63.7%)
Sentiment140	mean	3,528 ($\approx$ 2.4%)	6,254 ($\approx$ 4.2%)	138,930 ($\approx$ 93.4%)
Sentiment140	median	3,544($\approx$ 2.4%)	3,168 ($\approx$ 2.1%)	142,000 ($\approx$ 95.5%)

All the data were numeric, ranging from zero to tens of thousands. Thus, to convert this sparse data into meaningful machine-interpretable data, we applied a feature-scaling technique. We used a standardization methodology (Z-score normalization) whereby the features were rescaled so they would have the properties of a standard normal distribution. We built three kinds of machine learning models broadly classified as follows:

1. Classification models to predict the tweet sentiments as binary class labels (positive and negative)
2. Classification models to predict the tweet sentiments as one of three class labels (positive, neutral, and negative)
3. Regression models to predict the exact tweet sentiment scores.

As we built the machine learning models for a target variable that was purely

based on the sentiment analysis libraries, it was important to verify that the sentiment score generated using these libraries were reliable. For this purpose, two individuals manually labeled a random set of 200 tweets for positive and negative sentiment. For the manually labeled sentiments, we evaluated the Cohen's kappa coefficient (Cohen, 1960), which is a statistic for determining inter-rater reliability. The Cohen's kappa coefficient was 0.71. On validating the sentiments generated with the TextBlob library, we found that they matched with 84% and 81% of the two sets of manually labeled sentiments.

The tweets can contain the title of the research articles. To check the impact of the article's title in the tweets for tweet sentiment prediction, we performed the above-mentioned machine learning techniques with and without the tweets that had the title of the research articles. We performed word sequence matching on the tweets to check if they have the title of the research article and removed the tweets that had 70% or more sequence matches. We found that 105,834 articles had tweets that did not match with the title of the article. We further refer to the dataset with 148,712 articles (including tweets with article's title) as *dataset A* and the dataset with 105,834 articles (excluding tweets with article's title) as *dataset B*.

5. Results

5.1. Classification models

To create a model capable of predicting the binary sentiment as positive or negative, we decided to use NLTK VADER and TextBlob. We used these two libraries to obtain the sentiment score between -1 and 1 for the article title, abstract, and tweet (target variable). As the scores were in the range of -1 and 1, all the scores were

Table 5: Segregation of sentiments score.

Score range	Sentiment
[-1,0)	Negative
0	Neutral
(0,1]	Positive

transformed as positive, neutral, or negative sentiments (Table 5).

A research article may have multiple tweets related to it, and these tweets may differ from each other in terms of the sentiments expressed. For any given article, we calculated the mean and median of all the sentiments expressed in tweets about it. Table 6 shows some examples with tweets from the dataset to demonstrate the assignment of a sentiment label using the mean of tweets' sentiment.

A summary of the variations of the articles tweets' sentiments is shown in Tables 7 and 8. In all four cases mentioned, the sentiment library and the metric for multiple sentiments changed while the data was the same. Table 7 shows the variations of sentiments on dataset A and Table 8 shows the variations of sentiments on dataset B.

For classification, we have certain parameters such as

- True Negative (TN): Observation is negative and is predicted as negative.
- False Positive (FP): Observation is negative but is predicted as positive.
- False Negative (FN): Observation is positive but is predicted as negative.
- True Positive (TP): Observation is positive and is predicted as positive.

The evaluation metrics used for classification models are based on these four mentioned parameters:

Table 6: Examples of sentiment label assignment.

Article	**1st Tweet and Sentiment**	**2nd Tweet and Sentiment**	**3rd Tweet and Sentiment**	**Mean of tweets' sentiment**	**Final sentiment class label**
Article 1	Researchers in Norway investigate mortality risk of individuals after the death of a spouse (-0.7184)	Can you die of a broken heart? If your spouse dies, your death risk substantially increases (-0.9186)	A sad study: spouses much more likely to die after being widowed (-0.885)	-0.8407	Negative
Article 2	Presentation of the ABC Best Paper Award 2013 to Sherrie Elzey. Read the winning paper (0.9022)	ABC Best Paper Award 2013 goes to lead authors Sherrie Elzey and De-Hao Tsai. Read their article for free (0.9001)	NA	0.90115	Positive
Article 3	Latest article from our research team has been published about using School Function Assessment! (0)	Article on using School Function Assessment now online (0)	NA	0	Neutral

Table 7: Sentiments on dataset A using different libraries and metrics.

Experiment	Sentiment library	Metric for multiple sentiments	Number of positive sentiments	Number of negative sentiments	Number of neutral sentiments
case 1	VADER	mean	55,833 ($\approx$ 37.5%)	37,957 ($\approx$ 25.5%)	54,922 ($\approx$ 36.9%)
case 2	VADER	median	45,606 ($\approx$ 30.6%)	32,754 ($\approx$ 22%)	70,352 ($\approx$ 47.3%)
case 3	TextBlob	mean	67,035 ($\approx$ 45%)	16,881 ($\approx$ 11.3%)	64,796 ($\approx$ 43.6%)
case 4	TextBlob	median	53,466 ($\approx$ 36%)	13,748 ($\approx$ 9.2%)	81,498 ($\approx$ 54.8%)

Table 8: Sentiments on dataset B using different libraries and metrics.

Experiment	Sentiment library	Metric for multiple sentiments	Number of positive sentiments	Number of negative sentiments	Number of neutral sentiments
case 1	VADER	mean	44,866 ($\approx$ 42.4%)	26,664 ($\approx$ 25.1%)	34,304 ($\approx$ 32.4%)
case 2	VADER	median	38,038 ($\approx$ 35.9%)	23,124 ($\approx$ 21.8%)	44,672 ($\approx$ 42.2%)
case 3	TextBlob	mean	54,169 ($\approx$ 51.1%)	11,841 ($\approx$ 11.1%)	39,824 ($\approx$ 37.6%)
case 4	TextBlob	median	45,254 ($\approx$ 42.7%)	9,551 ($\approx$ 9%)	51,029 ($\approx$ 48.2%)

Accuracy shows the ratio of correct predictions to total observations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Precision is the ratio of the True Positive parameter to the sum of True Positive and False Positive parameters.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is the ratio of the True Positive parameter to the sum of True Positive

and False Negative parameters.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F-1 score is the weighted average of Precision and Recall and derives a balance between Precision and Recall. Unlike accuracy, the F-1 score considers both false positives and false negatives.

$$F\text{-}1\ score = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Weighted Average Precision/Recall/F-1 Score calculates the metrics of each label and finds the average weighted by support (the number of true instances for each label).

5.1.1. Classification models with two class labels

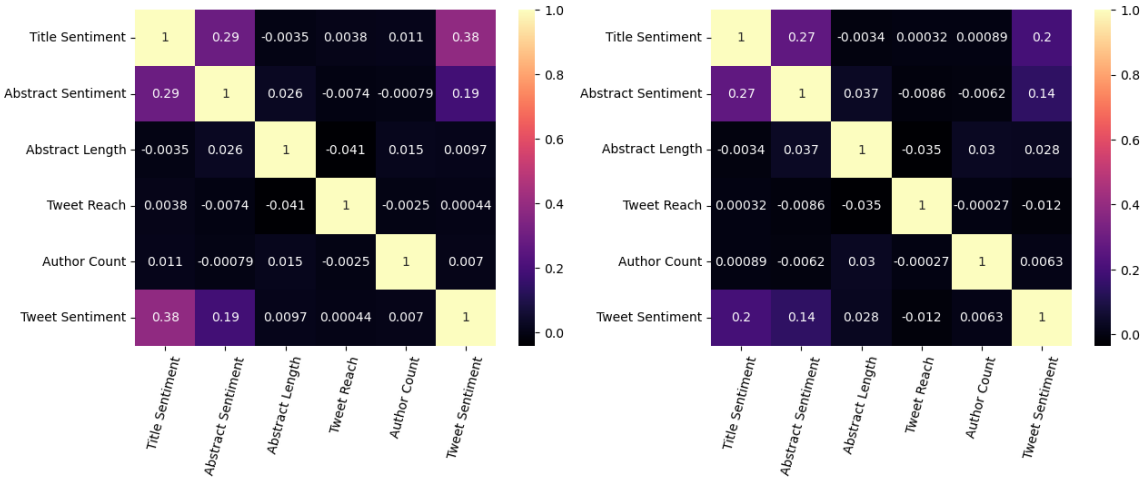
Here, we exclude neutral sentiments to consider only the positive and negative sentiments expressed in the tweets. The number of positive and negative tweets differs especially in cases 3 and 4. To overcome this class imbalance, we used the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002), which creates synthetic minority class examples. Additionally, we used 10-fold cross-validation and the Grid Search mechanism, which performs an exhaustive search over given parameters to build the best possible model (Pedregosa et al., 2011). We built machine learning models with an 80–20 train test split to predict the binary sentiment of the tweets for all four cases for both datasets A and B (Tables 7 and 8). In both scenarios, the case 4 model returned the highest accuracy.

Case 4 Experiment:

Here, we used TextBlob as the sentiment library, the median of the sentiments as the metric for multiple tweets to predict whether the tweet sentiments were positive or negative. Figure 4 shows the correlation matrix for the features used in this experiment for datasets A and B. On comparing the correlation of title sentiment with the sentiment of the tweets in Figures 4a and 4b, we observed a significant drop in the correlation on dataset B. Figure 5 shows a comparison of the performance of machine learning models in terms of accuracy and weighted average scores for precision, recall and F-1. The Random Forest and Decision Tree models performed better than the other models for both datasets A and B. From Figures 5a and 5b, we observed that the performance of Decision Tree and Random Forest models did not vary much, but there was a significant performance variation for Logistic Regression and K-Nearest Neighbors models. The important features for the binary classification derived from the Random Forest model are shown in Figure 6. The sentiment of the title of the research article is the most important feature, followed by the author count on both datasets A and B. From Figures 6a and 6b, we observe that irrespective of the title being in the tweet, the title's sentiment is still the most important feature for predicting the tweet sentiment. Table 9 shows the best results for cases 1–3 in predicting positive and negative tweet sentiments.

5.1.2. Classification models with three class labels

From Tables 7 and 8, we can observe that the neutral sentiments are numerous—a point that cannot be neglected in the analysis. For this experiment, the sentiment of a research article has three possible classes. Here, we also used SMOTE to deal with class-imbalance problems by upsampling the minority class label data. Additionally,



(a) Dataset A: Tweets with article's title.

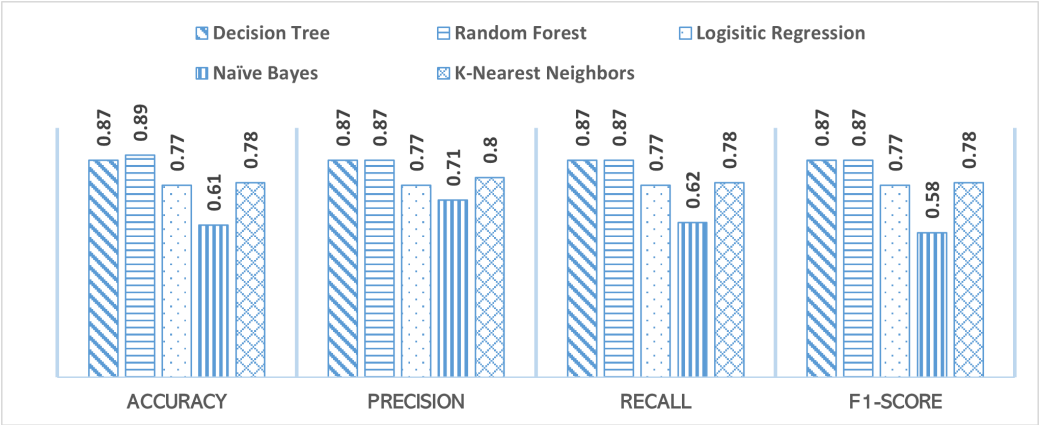
(b) Dataset B: Tweets without article's title.

Figure 4: Correlation matrix of features with two class labels – case 4.

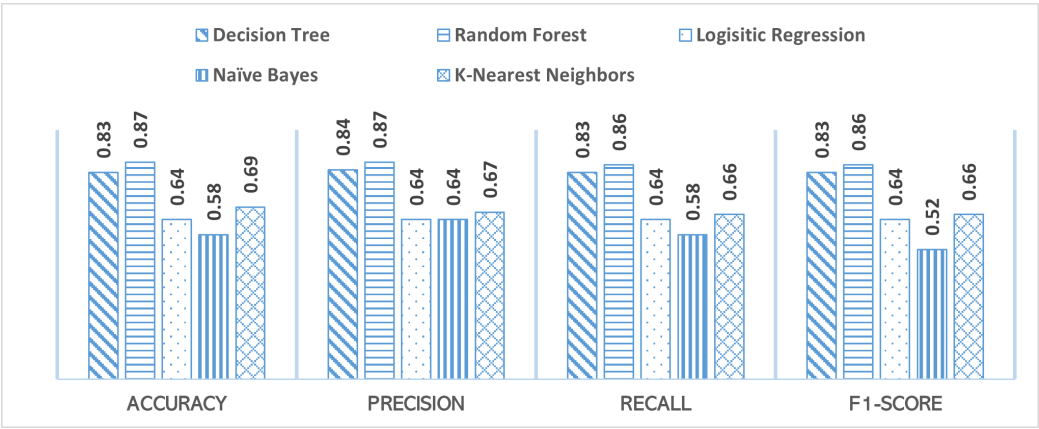
Table 9: Best results for cases 1–3 with two-class labels.

Dataset A: Tweets with article's titles			
Case Number	Model	Accuracy	F-1 Score
1	Random Forest	0.81	0.81
2	Random Forest	0.83	0.83
3	Random Forest	0.85	0.85
Dataset B: Tweets without article's titles			
Case Number	Model	Accuracy	F-1 Score
1	Random Forest	0.77	0.76
2	Random Forest	0.78	0.78
3	Random Forest	0.85	0.80

we used a 10-fold cross-validation and Grid Search mechanism to build the best possible model. We built machine learning models with an 80–20% train test split for all four cases on datasets A and B (Tables 7 and 8) that predict positive, neutral, and negative tweet sentiments. We observed that the case 4 experiment set-up generated better results than did the set-ups used for other cases on both datasets A and B.



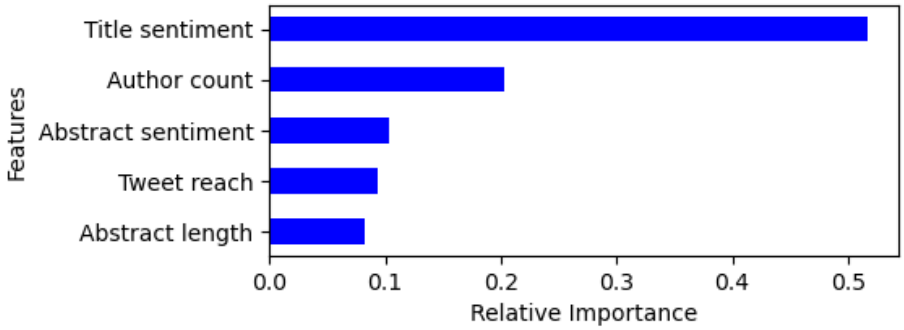
(a) Performance of models on Dataset A: Tweets with article's title.



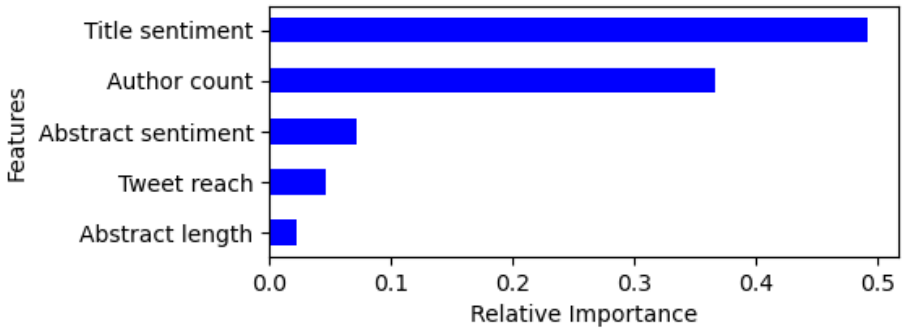
(b) Performance of models on Dataset B: Tweets without article's title.

Figure 5: Performance of classification models with two class labels – case 4.

Case 4 Experiment: Here, we used TextBlob as the sentiment library, the median of the sentiments as the metric for multiple tweets to predict whether the tweet sentiments were positive, neutral, or negative. Figure 7 shows the correlation matrix for the features used in this experiment for datasets A and B. We observed that the title sentiment of the research article had the highest correlation with the tweet sentiment. From Figures 7a and 7b, we observed that the correlation between title sentiment and tweet sentiment decreased by almost half from dataset A to B.



(a) Important features on Dataset A: Tweets with article's title.

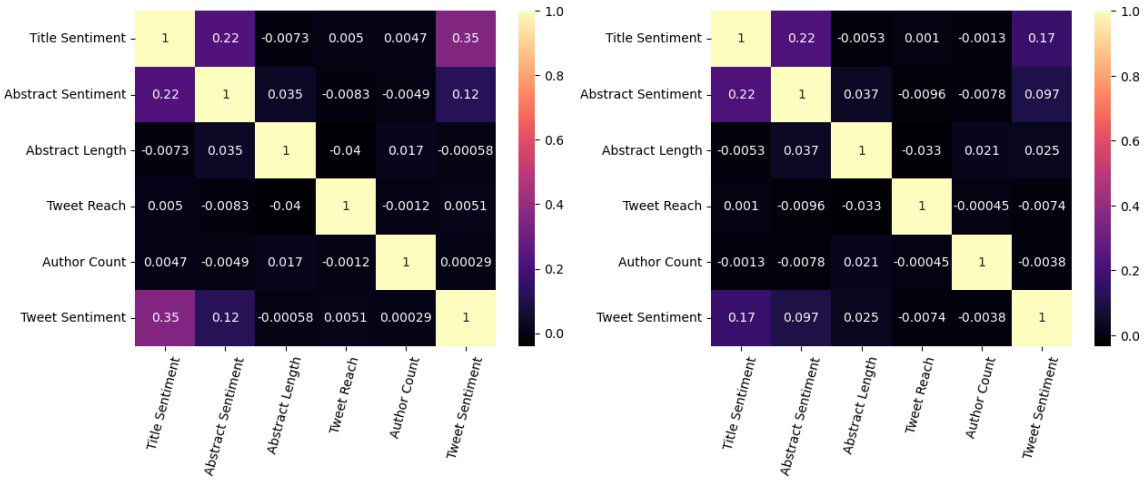


(b) Important features on Dataset B: Tweets without article's title.

Figure 6: Important features for two-class label classification.

Figure 8 shows the performance of the different models in terms of accuracy and weighted average scores for precision, recall, and F-1. Again, the tree-based models performed better than the other models on datasets A and B. Naive Bayes was the worst model. From Figures 8a and 8b, we observed that the performance of the models significantly dropped compared with the previous two class classification. Surprisingly, the Naive Bayes model performed better on dataset B. Figure 9 shows the important features derived from the Random Forest model. We observed that the sentiment of the research article's title is a vital feature in both datasets A and B. However, on comparing Figures 9a and 9b, we observed that the importance of

title sentiment decreases by about 15% when the model is built using the tweets that do not contain the article's title. Table 10 shows the best results for cases 1–3 in predicting positive, neutral, and negative tweet sentiments. Here, we observe that cases 1 and 2 that build the models on the VADER sentiment library did not exceed 50% for accuracy and F-1 score.



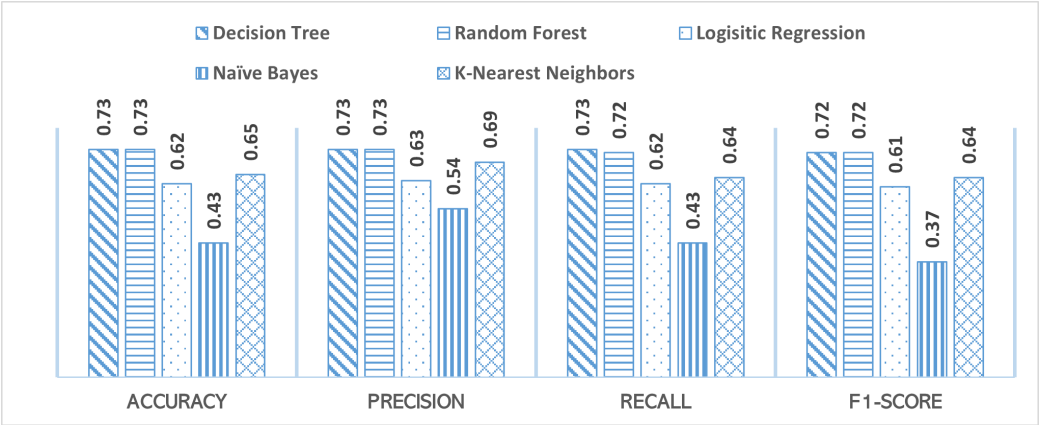
(a) Dataset A: Tweets with article's title.

(b) Dataset B: Tweets without article's title.

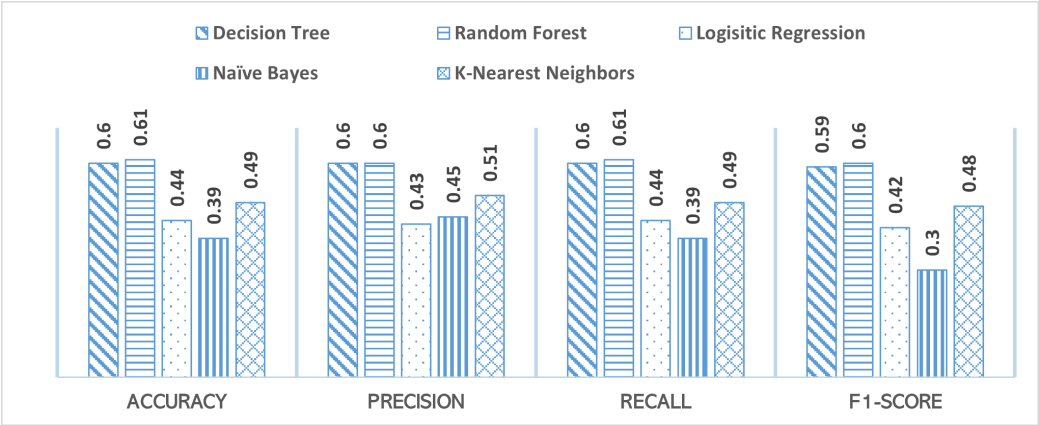
Figure 7: Correlation matrix of features with three class labels – case 4.

Table 10: Best results for cases 1–3 with three labels.

Dataset A: Tweets with article's titles			
Case Number	Model	Accuracy	F-1 Score
1	Random Forest	0.46	0.46
2	Random Forest	0.49	0.45
3	Random Forest	0.68	0.66
Dataset B: Tweets without article's titles			
Case Number	Model	Accuracy	F-1 Score
1	Random Forest	0.46	0.45
2	Random Forest	0.47	0.44
3	Random Forest	0.56	0.56



(a) Performance of models on Dataset A: Tweets with article's title.

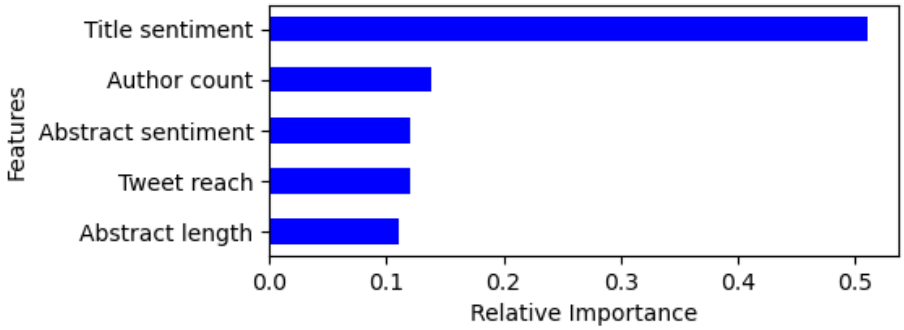


(b) Performance of models on Dataset B: Tweets without article's title.

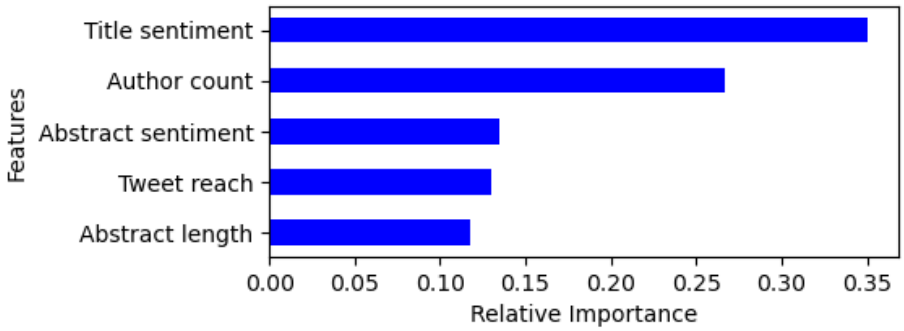
Figure 8: Performance of classification models with three class labels – case 4.

5.2. Regression models

We implemented some regression models to predict the sentiment score of the tweet. We used 80% training and 20% testing split and obtained the results shown in Table 11. The regression models did not perform well based on the R-squared values.



(a) Important features on Dataset A: Tweets with article's title.



(b) Important features on Dataset B: Tweets without article's title.

Figure 9: Important features for three-class label classification.

6. Discussion

In this study, to predict the tweet sentiments for research articles, we used the NLTK VADER and TextBlob libraries to obtain sentiment scores. We built machine learning models to predict the sentiment as a binary class label (positive or negative) and as three class labels (positive, neutral, or negative). We also used the Stanford CoreNLP library for a sample of 7,000 articles. However, this library resulted in about 95% of the sentiments being neutral. We observed that the dataset with sentiments generated using the TextBlob library had lower correlations among the title sentiment, abstract sentiment, and tweet sentiment when compared to those correlations on the

Table 11: Results of the regression models.

Dataset A: Tweets with article's titles		
Model	Mean Squared Error	R-Squared
Multiple Linear Regression	0.091	0.008
Decision Tree	0.189	-1.051
Random Forest	0.104	-0.130
Support Vector Regression	0.093	-0.014
Dataset B: Tweets without article's titles		
Model	Mean Squared Error	R-Squared
Multiple Linear Regression	0.104	0.006
Decision Tree	0.470	-1.095
Random Forest	0.119	-0.133
Support Vector Regression	0.106	-0.009

dataset having sentiments generated using NLTK VADER library.

Several studies that analyzed sentiments of tweets for scholarly articles have shown that the majority of the tweets for altmetrics have no sentiment value (neutral), and positive tweets are more than negative tweets. Thelwall et al. (2013) analyzed 270 tweets for papers published in 2012 and manually labeled 96% neutral, 4% positive, 0% negative. Friedrich et al. (2015b) performed sentiment analysis on 1,000 tweets using the SentiStrength tool and found that 94.8% of tweets were neutral, 4.3% positive, and 0.9% negative. Friedrich et al. (2015a) analyzed 487,610 tweets using SentiStrength and found 81.7% neutral, 11% positive, and 7.3% negative. Our study also shows that sentiments of tweets are more neutral than positive or negative. By using NLTK VADER and TextBlob libraries, we avoided the need to manually label tweets and have a lower level of neutral sentiments (43.6%), which reduces the class imbalance issue to an extent. Additionally, we used SMOTE to overcome class imbalance issues.

Any given research article could be discussed in multiple tweets. Therefore, we evaluated the final sentiment of all the tweets related to each given research article using the mean and median of the sentiments. However, we found that the mean and median did not matter much for the prediction of the tweet sentiment.

Raamkumar et al. (2018) studied tweets of computer science research articles using the Microsoft academic graph (MAG) dataset and found that the research impact (in terms of bibliometrics and altmetrics) of papers that had the three sentiments were better than those having just neutral sentiment. Interestingly, their study also used the TextBlob library to classify 49,849 tweets related to 12,967 computer science research papers. They found that 97.16% of the tweets had neutral sentiments, 2.8% had positive, and 0.05% had negative sentiments. Compared to our study, we observe that although the dataset is different, the library for sentiment classification is the same and tweets are still about the scholarly articles. The discrepancy in the percentage of the three sentiments could be because of the scientific field selection. Various fields might have various sentiment values for tweets. In computer science, users might just post the article with the given title without giving any reactions. However, most of our papers are health-related, which could explain the interest and increase in non-neutral sentiments.

In this study, we observed that the tree-based classifiers generally performed better than the other classifiers did. The best model we built was Random Forest using a median of tweets and the TextBlob library (case 4). When the tweets having the research article's title were included when building the model, it generated 89% accuracy for the binary classification and 73% accuracy for the three-label classification. When those tweets were excluded when building the model, it generated 73% accu-

racy for binary classification and 61% accuracy for the three-label classification. The most important feature in all the models was the sentiment of the research article's title.

We built regression models to determine the exact sentiment value for a tweet in a range of -1 (highly negative) to +1 (highly positive). However, these results were not as good as those from the classification with two or three class labels. Thus, we found it is difficult to determine the exact sentiment as a continuous value. However, in general, we were able to determine whether a tweet expressed a positive, negative, or neutral sentiment toward a given research article.

Our work has some limitations, such as considering several research disciplines that might affect the sentiments. For example, some disciplines might share similar keywords in the title and abstract but receive different sentiments. Additionally, we did not consider the impact of the publisher highlights on the tweet sentiment. Further, tweet sentiment may reflect a variety of factors but are not necessarily related directly to the specifics of the paper. The tweet sentiment, by no means, is an accurate indicator of the reaction to the research paper content. For example, we have no easy way of knowing if the tweeter read the study or just replied to the title or abstract. Furthermore, some users might have positive or negative sentiments toward scientific topics in general (e.g., pro- and anti-vaccination views) and not necessarily the scientific paper they are tweeting about. However, the sentiment of tweets is a significant metric for comprehending what the tweeters feel about the research. In a broader sense, this contributes to our understanding of the societal impact of research.

7. Conclusion and Future Work

In this study, we analyzed tweets related to research articles. We observed various patterns for the number of tweets, the research articles' subjects, and the publication years of the research articles. Due to the fact that tweets often include the title of the research paper, we created two distinct datasets: one with the title of the article in the tweets and one without. We built machine learning models to predict the sentiment of the tweet for a given research article. We developed classification models for two class and three class labels (positive, neutral, or negative). The tree-based classifiers generally outperformed other classifiers. We discovered that the sentiment expressed in the research paper title was the most significant feature in all models, followed by the number of authors. In future research, we plan to use neural networks to develop more robust machine learning models and provide tailored recommendations (Alhoori, 2016; Alhoori and Furuta, 2017) to authors. We also plan to expand the research area using word2vec models and extract important features from the title and the abstract of the research article that contribute to predicting the sentiments expressed in tweets.

Acknowledgments

This work is supported in part by NSF Grant No. 2022443.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pages 702–707. IEEE.

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., and Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, 15(2):101128.

Alhoori, H. (2016). How to identify specialized research communities related to a researcher's changing interests. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 239–240, New York, NY, USA. Association for Computing Machinery.

Alhoori, H. and Furuta, R. (2014). Do altmetrics follow the crowd or does the crowd follow altmetrics? In *IEEE/ACM Joint Conference on Digital Libraries*, pages 375–378.

Alhoori, H. and Furuta, R. (2017). Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics*, 11(2):553–563.

Alhoori, H., Furuta, R., Tabet, M., Samaka, M., and Fox, E. A. (2014). Altmetrics for country-level research assessment. In *International Conference on Asian Digital Libraries*, pages 59–64. Springer.

Alhoori, H., Ray Choudhury, S., Kanan, T., Fox, E., Furuta, R., and Giles, C. L. (2015). On the relationship between open access and altmetrics. *iConference 2015 Proceedings*.

Alhoori, H., Samaka, M., Furuta, R., and Fox, E. A. (2019). Anatomy of scholarly information behavior patterns in the wake of academic social media platforms. *International Journal on Digital Libraries*, 20(4):369–389.

Amolik, A., Jivane, N., Bhandari, M., and Venkatesan, M. (2016). Twitter sentiment analysis of movie reviews using machine learning techniques. *International Journal of Engineering and Technology*, 7(6):1–7.

Bae, Y. and Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and technology*, 63(12):2521–2535.

Bermingham, A. and Smeaton, A. (2011). On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10.

Bharathwaj, S. K., Na, J.-C., Sangeetha, B., and Sarathkumar, E. (2019). Sentiment analysis of tweets mentioning research articles in medicine and psychiatry disciplines. In *International Conference on Asian Digital Libraries*, pages 303–307. Springer.

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., and Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, X., Cho, Y., and Jang, S. Y. (2015). Crime prediction using twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium*, pages 63–68. IEEE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Da Silva, N. F., Hruschka, E. R., and Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179.

Didegah, F., Mejlgaard, N., and SÃ¸rensen, M. P. (2018). Investigating the quality of interactions and public engagement around scientific papers on twitter. *Journal of Informetrics*, 12(3):960–971.

Freeman, C., Alhoori, H., and Shahzad, M. (2020). Measuring the diversity of facebook reactions to research. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–17.

Freeman, C., Roy, M. K., Fattoruso, M., and Alhoori, H. (2019). Shared feelings: Understanding facebook reactions to scholarly articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 301–304. IEEE.

Friedrich, N., Bowman, T. D., and Haustein, S. (2015a). Do tweets to scientific articles contain positive or negative sentiments. In *Altmetrics Workshop, Amsterdam. Retrieved from http://altmetrics. org/altmetrics15/friedrich*.

Friedrich, N., Bowman, T. D., Stock, W. G., and Haustein, S. (2015b). Adapting sentiment analysis for tweets linking to scientific papers. *arXiv preprint arXiv:1507.01967*.

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94.

Hansson, K. and Ludwig, T. (2019). Crowd dynamics: Conflicts, contradictions, and community in crowdsourcing. *Computer Supported Cooperative Work (CSCW)*, 28(5):791–794.

Hansson, K., Ludwig, T., and Aitamurto, T. (2019). Capitalizing relationships: Modes of participation in crowdsourcing. *Computer Supported Cooperative Work (CSCW)*, 28(5):977–1000.

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L.-E., and Hsu, M.-C. (2011). Visual sentiment analysis on twitter data streams. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 277–278. IEEE.

Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In *2013 international conference on social computing*, pages 357–364. IEEE.

Hassan, S.-U., Saleem, A., Soroya, S. H., Safder, I., Iqbal, S., Jamil, S., Bukhari, F., Aljohani, N. R., and Nawaz, R. (2020). Sentiment analysis of tweets through altmetrics: A machine learning approach. *Journal of Information Science*, page 0165551520930917.

Haunschild, R., Leydesdorff, L., and Bornmann, L. (2020). Library and information science papers discussed on twitter: A new network-based approach for measuring public attention. *Journal of Data and Information Science*, 5(3):5.

Haustein, S. (2019). Scholarly twitter metrics. In *Springer handbook of science and technology indicators*, pages 729–760. Springer.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

Ibrahim, N. F. and Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers in Human Behavior*, 96:32–45.

Jaidka, K., Guntuku, S. C., Lee, J. H., Luo, Z., Buffone, A., and Ungar, L. H. (2021). The rural–urban stress divide: obtaining geographical insights through twitter. *Computers in Human Behavior*, 114:106544.

Kale, B., Siravuri, H. V., Alhoori, H., and Papka, M. E. (2017). Predicting research that will be cited in policy documents. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 389–390, New York, NY, USA. Association for Computing Machinery.

Kharde, V., Sonawane, P., et al. (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

Kou, Y., Kow, Y. M., Gui, X., and Cheng, W. (2017). One social movement, two social media sites: A comparative study of public discourses. *Computer Supported Cooperative Work (CSCW)*, 26(4):807–836.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Kousha, K. and Thelwall, M. (2019). An automatic method to identify citations

to journals in news stories: A case study of uk newspapers citing web of science journals. *Journal of Data and Information Science*, 4(3):73–95.

Le, X., Chu, J., Deng, S., Jiao, Q., Pei, J., Zhu, L., and Yao, J. (2019). Citeopinion: Evidence-based evaluation tool for academic contributions of research papers based on citing sentences. *Journal of Data and Information Science*, 4(4):26.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., and Rodrigue, J. (2012). A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the second workshop on language in social media*, pages 27–36.

McKinney, W. et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Mittal, A. and Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15.

Mohammad, S. M. (2017). Challenges in sentiment analysis. In *A practical guide to sentiment analysis*, pages 61–83. Springer.

Narr, S., Hulfenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12–14.

Neethu, M. and Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Noyons, E. (2019). Measuring societal impact is as complex as abc. *Journal of data and information science*, 4(3):6–21.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of European Language Resource Association*, volume 10, pages 1320–1326.

Pal, J., Thawani, U., Van Der Vlugt, E., Out, W., Chandra, P., et al. (2018). Speaking their mind: Populist style and antagonistic messaging in the tweets of donald trump, narendra modi, nigel farage, and geert wilders. *Computer Supported Cooperative Work (CSCW)*, 27(3):293–326.

Pandian, N. D. S., Na, J.-C., Veeramachaneni, B., and Boothaladinni, R. V. (2019). Altmetrics: Factor analysis for assessing the popularity of research articles on twitter. *Journal of Information Science Theory and Practice*, 7(4):33–44.

Parikh, R. and Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. *CS224N Final Report*, 118.

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

Pozzi, F. A., Fersini, E., Messina, E., and Liu, B. (2017). Challenges of sentiment analysis in social networks: an overview. *Sentiment analysis in social networks*, pages 1–11.

Raamkumar, A. S., Ganesan, S., Jothiramalingam, K., Selva, M. K., Erdt, M., and Theng, Y.-L. (2018). Investigating the characteristics and research impact of sentiments in tweets with links to computer science research papers. In *International Conference on Asian Digital Libraries*, pages 71–82. Springer.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shaban, H. (2019). Twitter reveals its daily active user numbers for the first time. *Washington Post*, 7.

Shaikh, A. R. and Alhoori, H. (2019). Predicting patent citations to measure economic impact of scholarly research. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 400–401.

Siravuri, H. V., Akella, A. P., Bailey, C., and Alhoori, H. (2018). Using social media and scholarly text to predict public understanding of science. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 385–386, New York, NY, USA. Association for Computing Machinery.

Siravuri, H. V. and Alhoori, H. (2017). What makes a research article newsworthy?

Proceedings of the Association for Information Science and Technology, 54(1):802–803.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558.

Thelwall, M., Tsou, A., Holmberg, K., and Haustein, S. (2013). Tweeting links to academic articles. *Cybermetrics*, 17(1):1–8.

Vinkers, C. H., Tijdink, J. K., and Otte, W. M. (2015). Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: retrospective analysis. *Bmj*, 351.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012a). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120.

Wang, X., Gerber, M. S., and Brown, D. E. (2012b). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. (2011). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1031–1040.

Zaman, T. R., Herbrich, R., Van Gael, J., and Stern, D. (2010). Predicting infor-

mation spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, pages 17599–601. Citeseer.