EPiC Series in Computing
Volume 98, 2024, Pages 25–35
Proceedings of 39th International Conference on Computers and Their Applications

Exploring Sentiment on Campus. A Twitter Sentiment Analysis on University Tweets

Alina Campan, Murtadha Almakki, Trang Do, and Traian Marius Truta
School of Computing and Analytics
Northern Kentucky University, Highland Heights, Kentucky, U.S.A.
campana1@nku.edu, almakkim2@mymail.nku.edu, dot2@mymail.nku.edu, trutat1@nku.edu

Abstract

We report in this work the results of our analysis of accuracy of 5 sentiment analysis methods (TextBlob, VADER, logistic regression, support vector machine, CNN on encodings based on BERT tokenization,) for a dataset consisting of tweets from the academia domain, that we API-scraped for 32 universities during the year 2022. We show some results for the volume and sentiment polarity trends exhibited by this dataset. We connect peak and low sentiment averages to concrete events that explain the respective sentiment trend; this proves that observing the social media trends allows to detect real events that need attention and possible action.

1 Introduction

Online social platforms are extensively used and play a significant role in the daily life of many individuals. A variety of social media platforms exist, with Facebook and Twitter (now X; for consistency we are referring to X as Twitter in this paper) being some of the most well-known. While Facebook has almost 3 billion users (2.958 billion monthly active users as of January 2023), Twitter is significantly smaller, having “only” 556 million users (as of January 2023) [1]. These platforms’ usage and privacy expectations are also different. On Facebook, the information posted is usually intended to a limited number of friends or groups. On Twitter, which is a micro-blogging platform, the goal of most users is for their posted messages to reach a wide audience. Therefore, there are different privacy concerns for the two platforms: Facebook is expected to ensure strict privacy for its users’ data, while Twitter users have fewer privacy concerns, as tweets are vehicles to expressing opinions that reach a large audience. The tight privacy requirements make data collection from Facebook difficult, technically, and, recently, data scraping is not permitted as per Facebook’s Terms of Service policy [2]. However, illegal scraping of Facebook data happened on a very large scale as recently as 2017, when Cambridge Analytica collected personal information of at least 87 million people, most of them from the United States [3]. To address this unauthorized collection, Facebook increased the restrictions on data access even more [4]. Twitter, on the other hand, allowed (until March 2023) tweets to be collected in large quantities, for research and business purposes, and offered APIs (application programming interfaces) through which programs could connect to the live Twitter stream and collect a sample [5]. Starting in March 2023, Twitter API v2 stopped free access to sample from the data stream, and the future of academic access to the platform, if any, is still unknown at this time (Fall 2023). The reasons to the restricted access are not privacy-related; instead, the restrictions are driven by an attempt to monetize the access to the Twitter data stream. Our collection of university-related tweets was conducted during calendar year 2022, through the Free Filtering API, which was available at the time, and we present the results of our analysis on the collected dataset. As Twitter might remain unavailable as a reasonably priced data source for academic research, the analysis we present here can be replicated similarly for whichever content source will be available in the future (Reddit, Digg, YouTube, Mastodon Social, BlueSky Social etc.)

Social media data has been analyzed for various application areas: effect of fake news, misinformation, and disinformation on election results [6] or vaccination hesitancy [7], observing detecting spread of viral diseases such the flu or covid-19 [8].

Recently, there has been work done in the academic and data analytics community as well, on opinion mining of content originating from online social media platforms, to determine positive or negative opinions towards selected institutions [9]. Such an analysis makes sense and can support decision making in academia, considering how social media is employed in the higher ed environment.

Social media is currently used by universities, colleges, departments, administrators, for news distribution and marketing. It is also employed in teaching and learning; some teachers use social media platforms as a tool to motivate, engage, and encourage student participation [9].

Students consider university rankings and opinions of their peers when choosing a school. Multiple well-known or reputable rankings of universities exist, such as Shanghai ARWU (Academic Ranking of World Universities) [10], THE (Times Higher Education) World University Rankings [11], and the U.S News Best National University Rankings [12]. The procedures to create these rankings must meet methodical standards, and they use various indicators related to the core missions of universities (such as teaching, research, knowledge transfer) or other aspects (such as reputation) [9]. These rankings, and even the procedures used in creating them, were sometimes criticized for reasons related to data collection, the weighting applied to various factors in creating the overall ranking measure, and because they measure institutions as a whole, and not smaller units like schools and departments [9]. Therefore, alternative or complementary rankings created by analyzing social-media content might be worth the effort. These rankings, that universities could create in-house by running data collection and analysis processes, could be more time-sensitive, more up-to-date, and meeting the desired granularity (e.g. not the whole institution, but specific units, like colleges, departments, and majors). By detecting real-time trends, universities could “cash in” on positive opinions, could work to create and increase good reputation, and could react quickly to address negative opinions generated by current events. In regards of students choosing the schools they want to attend, a question that is worth studying is how much impact official rankings and these unofficial, peers’ opinions might have on their decisions. Of course, the rankings we are referring to above are focusing primarily on general public perception regarding these universities, and not on exact quantitative measures as the ones used in currently existing rankings [10, 11, 12].

The contributions of this work are outlined below:

1. We present our methodology for collecting and analyzing the sentiment polarity in a university tweet dataset. Our API scraping process was run for the entire 2022 year, and we collected a total of 8,148,594 university related tweets.
2. We perform a comparative analysis of the accuracy with which different sentiment analysis (SA) methods classify data into several predefined groups (positive/negative/neutral.)

3. We present information about the collected tweets and the overall volume and sentiment trends for the entire tweet dataset for all 32 observed universities, for the entire time window, January-December 2022.

The structure of the paper is as follows. Section 2 presents our framework for data collection and processing methodology, and initial statistics on the collected tweets. Section 3 describes our comparative analysis of the accuracy of several sentimental analysis methods. Section 4 presents conclusions and future work plans.

2 University Dataset Collection and Processing

Our data collection process utilized the Twitter Free Streaming API with filtering, facilitated by the Python library Tweepy [13]. This API allowed to collect real-time streaming tweets that matched predefined keywords, usernames, and other filtering criteria. The workflow of our experiments consisted of the following steps:

1. **Data Collection**: Leveraging the Twitter Free Streaming API, we continuously collected tweets from the live stream throughout the year 2022. The search keywords for each university were defined based on hashtags and relevant terms. Example: for University of Cincinnati, we filtered by search words: #Bearcats, #NextLivesHere, UC Cincinnati, University Cincinnati, Bearcats, uofcincy, Prez_Pinto, GoBEARCATS.
2. **Data Preprocessing**: The collected tweets were sorted by their creation time to establish a chronological order. Duplicate tweets were removed to ensure data accuracy.
3. **Tweet Count Analysis**: We counted the total number of collected tweets and further analyzed the tweet volume within various time intervals, such as hours, days, and months.
4. **Sentiment Analysis**: To estimate the sentiment of the collected tweets, we employed a lexicon-based technique called VADER (Valence Aware Dictionary and sEntiment Reasoner). This technique assigns sentiment scores to text based on pre-defined sentiment scores of individual words.
5. **Sentiment Summarization**: We computed a moving cumulative average of sentiment scores for different time intervals (hours, days) to understand the overall sentiment trend over time.
6. **Volume and sentiment analysis for subgroups of tweets**: We designed a filtering method similar to Twitter's matching mechanism, to split the overall dataset into subgroups matching keywords for each of the observed 32 universities (overlaps allowed, as a tweet could match several universities' associated keywords.) We then counted and estimated sentiment for subsets of tweets (for each university,) similar to the process detailed above in steps 3-5 for the entire dataset.

Below, we detail how the data for our study was collected. Our research focuses on analyzing tweets related to 32 universities in the United States, categorized into different groups. The data collection process involved the use of hashtags and search terms to identify relevant tweets from the Twitter live stream. The universities were grouped into the following categories:

1. **KY Public Universities**: Eight public universities located in Kentucky.
2. **Tri-State Universities**: Universities within a 50-kilometer radius from our institution, Northern Kentucky University (NKU).
3. **KY Benchmark Universities**: Subset of KY Public Universities chosen as benchmarks.

4. **Horizon League Universities**: Universities belonging to the Horizon League conference, including NKU.
5. **Top US Universities**: Top 10 ranked US universities based on the U.S. News & World Report rankings.

The data collection strategy was customized to each university's distinct identity, employing specific hashtags and search terms to capture relevant tweets. Here are examples illustrating the hashtags and search terms for a few universities:

Northern Kentucky University (Tri-State Universities):

- *Hashtags:* #NorseUp, #Norsebound
- *Search Terms:* NKU Kentucky, Northern Kentucky University, @NKUEDU, @PrezVaidya, @NKUNorse

University of Cincinnati (Tri-State Universities):

- *Hashtags:* #Bearcats, #NextLivesHere
- *Search Terms:* UC Cincinnati, @uofcincy, @GoBEARCATS

University of Kentucky (Top US Universities):

- *Hashtags:* #WeAreUK, #ForTheTeam, #UKWildlyPossible
- *Search Terms:* Kentucky Wildcats, UK Kentucky, @universityofky

Harvard University (Top US Universities):

- *Hashtags:* #GoCrimson, #OneCrimson
- *Search Terms:* Harvard University, @Harvard, @harvardcrimson

The API scraping for university tweets was run for the entire year 2022, and the total number of collected tweets was: 8,148,594 .

Figure 1 illustrates the total count of university-related tweets collected throughout the study for each day. Significantly, there was an average of around 20,000 tweets per day, with nine days experiencing tweet volumes exceeding 50,000. Notably, on two particular days, May 20 and October 28, tweet volumes reached exceptionally high levels, surpassing 100,000 tweets. The peak occurred on October 28, with a volume of 182,475 tweets.

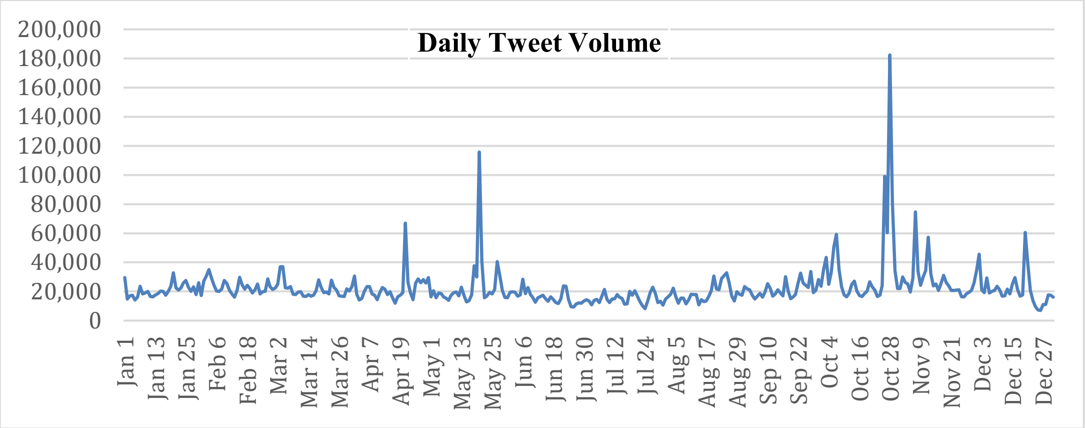

Figure 1: Number of University tweets collected daily during 2022

As discussed later in the paper (see Section 3), we found that the VADER (Valence Aware Dictionary for sEntiment Reasoning) model [14] for estimating sentiment polarity is reasonably accurate, while having the advantage of not requiring a training phase. Therefore, we used VADER to estimate the overall sentiment for the entire university-related dataset, for various time granularities (per second, per hour, per day etc.). In Figure 2, we showcase the average sentiment value per day. The calculated average sentiment score across all collected tweets was 0.15. Two days stood out with significant positive sentiment spikes (above 0.4), occurring on November 12 and December 2. Conversely, two days experienced negative sentiment lows (below -0.1), on August 20 and November 7. The lowest sentiment average of -0.157 occurred on November 7.

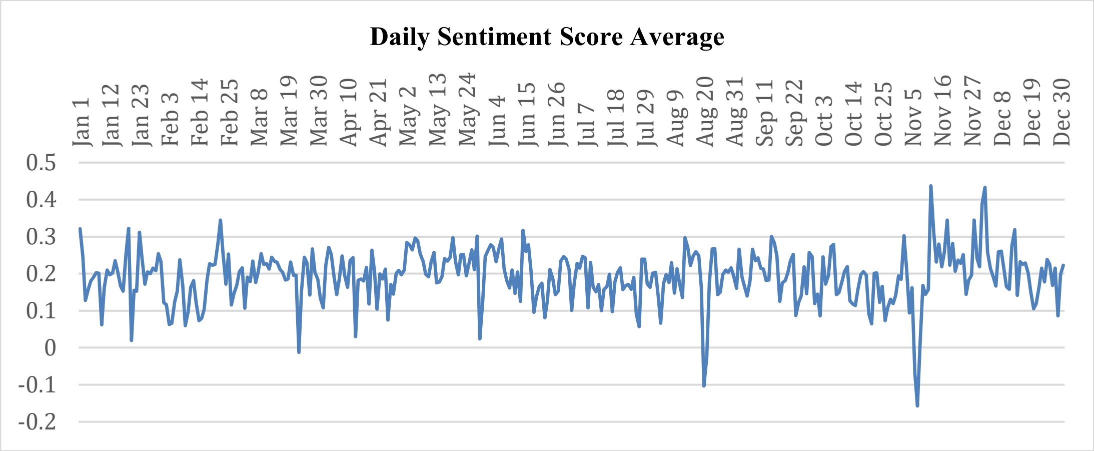

Figure 2: Average sentiment score for each day

Although the comprehensive examination of the factors contributing to the emergence of above-mentioned peaks is not within the scope of this paper, it is noteworthy that such peaks originate from distinct topics or events disseminated extensively through retweets and quoted tweets. For instance, on October 28, a viral and humorous tweet circulated, positing that scientists from Princeton University had reconstructed a 3D representation of how Adam might have appeared, accompanied by an image of the actor Vin Diesel. This particular tweet, along with its subsequent retweets, quotes, and imitations, engendered the most substantial surge in volume within our dataset for the entirety of the year 2022.

3 Sentiment Analysis Methods Comparisons

We performed a comparative analysis of the accuracy with which different sentiment analysis (SA) methods classify data into three predefined groups: positive, negative, and neutral. We trained several models (for the supervised methods) with training data obtained from multiple sources, unrelated to the university tweet dataset we collected. We tested each of the SA models with a test dataset sampled from our collected Twitter university dataset. Given the lack of relationship between our training and testing data (they are from different application domains,) it follows that the trained models are not overfit to our university tweet dataset; therefore, the accuracy of these trained models (and that of the rule-based systems as well) is not overestimated because the train and test datasets are not drawn from the same data distribution.

The training dataset. For training, we combined data from three sources.

1. 5,000 positive tweets and 5,000 negative tweets were taken from nltk's twitter_samples dataset [15], the positive_tweets.json and negative_tweets.json files; the sentiment class of a tweet is determined by the file that contains it.
2. 16,399 texts (tweets, MySpace messages, YouTube comments, Digg content) were taken from the Sentiment Strength Twitter Dataset (SS-Tweet [16]; in this case, the sentiment class of a text – positive, negative, or neutral, was decided by calculating the difference between the text's positive and negative sentiment strengths, and the difference being > 1 (for positive), < 0 (for negative), or in the interval [0, 1] (for neutral).
3. Finally, a sample of 8,000 movie reviews was taken from the IMDB Dataset [17] 4,003 of these reviews were positive and 3,997 were negative; the dataset already contains the sentiment class label for each review.

Overall, our training dataset had a total of 11,022 positive texts, 12,066 negative texts, and 5,655 neutral texts. For training logistic regression (LR) and support vector machine (SVM) models, we sampled 5,500 texts of each class, so the classes were balanced.

The test dataset. The test dataset is based on 730 tweets from our university-related dataset. We sampled randomly 730 tweets from this collection and manually labeled them as neutral, negative, or positive. The majority of the tweets in the test dataset were in the neutral category (500), 102 were negative, and 128 were positive.

In the realm of **sentiment analysis (SA) methods**, the terms "opinion mining" and "sentiment analysis" emerged concurrently in the early 2000s [18]. The nomenclature "opinion mining" found its roots in the web search and information retrieval communities [19], while "sentiment analysis" is more prevalent within the natural language processing (NLP) communities [20]. In a technical sense, sentiment analysis, also known as opinion mining, constitutes a computational procedure whereby a given text is categorized as either positive, negative, or neutral in sentiment. Furthermore, the outcome of a sentiment analysis can manifest as a continuous numerical value, typically ranging from -1 to 1. A smaller numerical value denotes a more negative sentiment. This numerical sentiment "value" within the [-1, 1] range is commonly referred to as polarity.

Numerous techniques are available to compute the sentiment of a given text. These methods fall into two primary categories: machine-learning approaches, such as classification, and lexicon-based approaches, which derive the meaning of a text from scores associated with individual words or textual fragments.

Two lexicon-based approaches (non-supervised methods) capable of determining the polarity of a tweet (or text in general) are **TextBlob** [21] and **VADER** [14]. TextBlob is a Python library designed for text data processing. It offers a user-friendly API for performing various NLP tasks, including part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more [21]. VADER (Valence Aware Dictionary for sEntiment Reasoning) employs a combination of qualitative and quantitative methods to generate a sentiment lexicon that is particularly well-suited for microblog-like contexts, making it optimized for social media analysis [14]. Both TextBlob and VADER are open-source tools that rely on annotated datasets, which are not specific to any particular domain, and can be directly employed to obtain sentiment scores without requiring prior data preparation.

Alternatively, supervised learning methods, specifically classification techniques, enable the training of models using domain-specific data. These trained models can subsequently be utilized to make predictions, either assigning a sentiment class (positive, negative, or neutral) or generating a numerical score for a new tweet that the model has never encountered before. In our research, we conducted experiments using two such supervised learning approaches for sentiment analysis: **logistic regression (LR)** [22] and **support vector machines (SVM)** [23]. A common prerequisite for all these methods is the initial training of a model using annotated data. The training dataset typically consists of tweets, with each tweet manually labeled by a researcher according to one of the predefined sentiment categories of interest. The learning algorithm then constructs a model by studying this annotated dataset, and once trained, the model can be used to provide predictions, either classifying new texts or assigning sentiment scores to them.

As our final method, we employed a BERT tokenizer to encode our texts and trained a model consisting of three convolutional neural network (CNN) layers using the BERT encodings [24]. This CNN model was trained on the entirety of the training dataset.

In Table 1, we show the training information, the thresholds for discretizing prediction scores into class labels, the accuracy score and F1-score values for the five sentiment analysis models described above, when evaluated on the University test dataset. To compute the accuracy and F1-scores we used the Scikit-learn library in python [25]. As a reminder, the University test dataset has 128 positive tweets, 102 negative tweets, and 500 neutral tweets.

	TextBlob	**VADER**	**3-class LR model**	**3-class SVM model**	**3-class CNN using BERT tokenizer encodings**
Training	None	None	Sample of 5,500 texts of each class from the training dataset. 5 models were trained and applied.	Sample of 5,500 texts of each class from the training dataset. 5 models were trained and applied.	Entire training set used to train the CNN.
Thresholds for discretizing prediction scores into class labels	score > 0 => 'Positive' class score < 0 => 'Negative' class score = 0 => 'Neutral' class	score > 0.05 => 'Positive' class score < -0.05 => 'Negative' class score = 0 => 'Neutral' class	None (Class with the highest probability is predicted)	None (Prediction is a class label, not a probability)	None (Class with the highest probability is predicted). The model classifies 688 out of 750 texts (42 can't be classified because text is not in English, or is too short)
Accuracy score	55.479	55.890	69.6164 (average for 5 models)	66.7394 (average for 5 models)	39.589
Macro F1-score	49.629	54.790	59.616 (average for 5 models)	58.9028 (average for 5 models)	30.614

Table 1: Accuracy and F1 values for SA models on the University test dataset with 3-classes

The confusion matrices for predictions of the five SA methods are shown in Figure 3. It can be seen that TextBlob miscategorizes especially the negative tweets, and incorrectly predicts a lot of content as belonging to the neutral category. VADER and LR mostly miscategorize the negative tweets – which means that the averages sentiment values reported will tend to be overestimates of the real sentiment trend of the tweet dataset (i.e. estimates are more positive than in reality.) Therefore, especially low sentiment estimates should be trusted as being reflective of the real sentiment. LR and SVM have higher errors on the positive and negative tweets prediction, compared with VADER and TextBlob, which mostly miscategorize the neutral tweets. Please note that for both LR and SVM, we report the confusion matrix for the model (out of 5 models we run for each) that has the highest accuracy and F1-score.

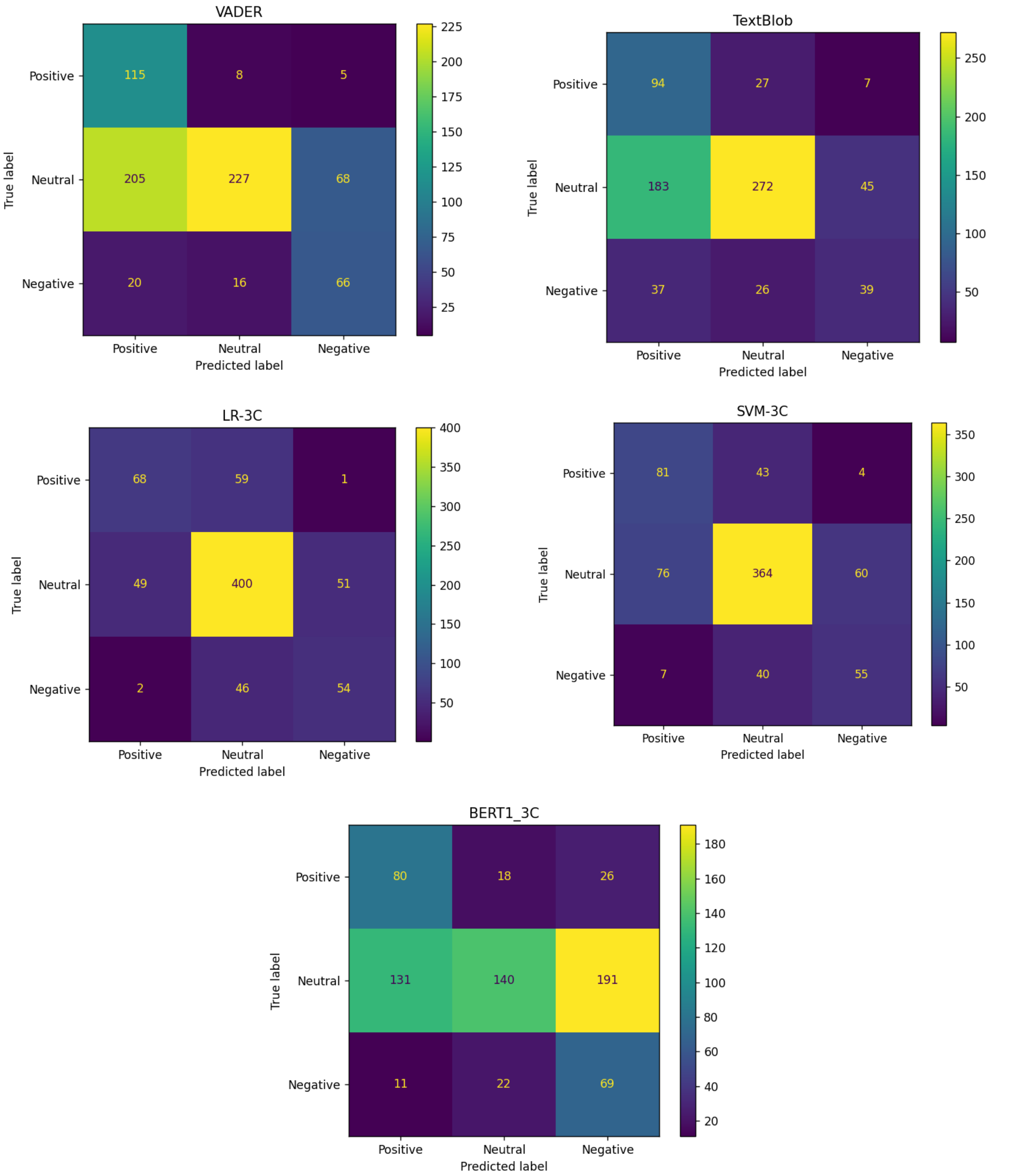

Figure 3: Confusion matrices for SA models on University test dataset

For comparison, we also report in Figure 4 the accuracy, F1-score, and confusion matrices of VADER and TextBlob when applied on the training set described in the beginning of this section. As mentioned above, VADER and TextBlob mostly miscategorize the neutral tweets, so the accuracy and F1-score are higher on the training set than on the test set, because of their composition: while the training set contains only 19.67% neutral tweets, the test set has a much higher percentage of neutral tweets (68.49%,) which are especially prone to miscategorization.

VADER

Accuracy: 65.696, *Macro F1-score*: 61.037

TextBlob

Accuracy: 70.208, *Macro F1-score*: 65.807

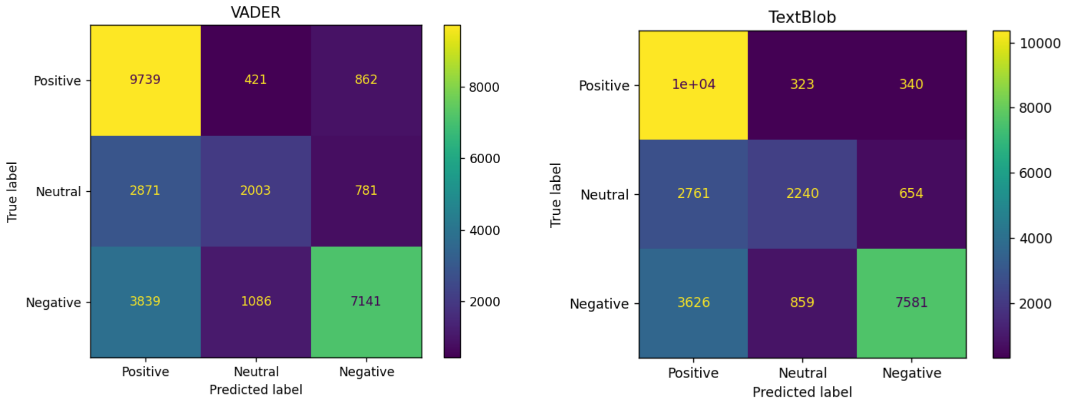

Figure 4: VADER and TextBlob accuracy, F1-score, and confusion matrix for the training set

4 Conclusions and Future Work

We reported in this work the results of our analysis of accuracy of 5 sentiment analysis methods, for a dataset consisting of tweets from the academia domain. We also showed some preliminary results for the volume and sentiment polarity trends exhibited by this dataset, for the year of 2022. We connected peak and low sentiment averages to concrete events that explain the respective sentiment trend; this proves that observing the social media trends allows to detect real events that need attention and possible action.

Our next steps will be to finalize and report an in-depth analysis of volume and sentiment for individual universities, or groups or universities (for example, KY benchmark universities.) We plan to add sentiment classification using a tuned BERT model, and determine if it allows for more accuract sentiment class prediction. We also plan to run topic analysis on the overall dataset and tweet subsets for individual universities/groups of universities – this will allow to *automatically* find most important topics discussed, including detecting the events that need action (in contrast, we had to manually identify the messages and topic driving the peaks and lows as reported in section 2.)

References

[1] Statista Global Social Networks, "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," 2023. Available: https://www.statista.com/statistics/272014/ global-social-networks-ranked-by-number-of-users.

[2] Facebook Terms of Service, "Terms of service," 2023. Available: https://m.facebook.com/terms.php.

[3] D. O'Sullivan, D. Griffin, P. DiCarlo, "Cambridge Analytica's Facebook data was accessed from Russia, MP says," in *CNNTech*, 2018. Available: https://money.cnn.com/2018/07/17/technology/ cambridge-analytica-data-facebook-russia/index.html.

[4] M. Schroephfer, "An update on our plans to restrict data access on Facebook," 2018. Available: https://newsroom.fb.com/news/2018/04/restricting-data-access.

[5] J. Littman, "Where to get Twitter data for academic research," 2017. Available: https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data.

[6] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 31 (2): 211-36, 2017.

[7] I. Montagni, K. Ouazzani-Touhami, A. Mebarki, N. Texier, S. Schück, C. Tzourio, and the CONFINS group, "Acceptance of a Covid-19 vaccine is associated with ability to detect fake news and health literacy," *Journal of Public Health*, 43 (4): 695–702, 2021.

[8] O.Shahid, M.Nasajpour, S.Pouriyeh, R.M. Parizi, M.Han, M.Valero, F. Li, M. Aledhari, Q. Z. Sheng, "Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance," *Journal of Biomedical Informatics*, Vol. 117, 2021.

[9] A. Abdelrazeq A., D. Janssen, C. Tummel, S. Jeschke, and A. Richert, "Sentiment analysis of social media for evaluating universities," in *Proceedings of the 2nd Intl. Conference on Digital Information Processing, Data Mining, & Wireless Communications (DIPDMWC 2015)*, Dubai, UAE, 2015.

[10] Shanghai Ranking, "Academic ranking of world universities (ARWU)," 2023. Available: https://www.shanghairanking.com.

[11] Times Higher Education, "Times Higer Education (THE) world university rankings," 2023. Available: https://www.timeshighereducation. com/world-university-rankings.

[12] U.S. News & World Report, "Best national university ranking," 2023. Available: https://www.usnews.com/best-colleges/rankings/national-universities.

[13] Tweepy, "An easy-to-use Python library for accessing the Twitter API," Accessed 2022. Available: https://www.tweepy.org.

[14] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*. 8 (1), 2014.

[15] S. Bird, E. Klein, E. Loper, "Natural language processing with Python: analyzing text with the natural language toolkit," *O'Reilly Media, Inc.*, 2009.

[16] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology*, 63 (1), 163–173, 2012.

[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[18] B. Liu, “Sentiment analysis and opinion mining. Synthesis lectures on human language technologies,” *Springer*, 2012.

[19] T. Nasukawa and J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” in *Proceedings of the 2nd International Conference on Knowledge Capture*, 70–77, 2003. Available: http://dx.doi.org/10.1145/945645.945658.

[20] K. Dave, S. Lawrence, D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in *Proceedings of the World Wide Web Conference (WWW)*, 2003.

[21] S. Loria, “Textblob documentation, Release 0.16.0,” 2020. Available: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf.

[22] J. S. Cramer, “The origins of logistic regression_(PDF) (Technical report),” Vol. 119, *Tinbergen Institute*. 167–178, 2002. Available: doi:10.2139/ssrn.360300.

[23] C. Cortes, and V. Vapnik, "Support-vector networks,", in *Machine Learning*, 20 (3), 273–97, 1995. Available: CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018. S2CID 206787478.

[24] U. Malik, “Text classification with BERT tokenizer and TF 2.0 in Python”. Available: https://stackabuse.com/text-classification-with-bert-tokenizer-and-tf-2-0-in-python.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, “Scikit-learn: Machine learning in Python,” in *The Journal of Machine Learning Research*, 12, 2825–2830. 2011.