Sentiment Analysis On Reddit Data And Analyzing Attacks on Social media

Shahana Sherfudeen (SS)
Sahithi NallaniChakravartula (SNC)

Abstract

In this project we have developed a model that classifies user reviews from real-time reddit data into positive, neutral and negative categories using ML models. We have analyzed the Attack and defense methods through 4 reference papers and given the insights through literature review. The growing use of social media has led to the development of several Machine Learning (ML) and Natural Language Processing (NLP) tools to process the unprecedented amount of social media content to make actionable decisions. Sentiment Analysis is the process of computationally identifying and categorizing words expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive or negative. Any brand's presence on social networks has a significant impact on emotional reactions of its users to different types of posts on social media. If a company understands the preferred types of posts (photo or video) of its customers, based on their reactions, it could make use of these preferences in designing its future communication strategy. In sentiment analysis, an attacker can launch adversarial attacks by adding small perturbations to text to generate different perceptions than the actual opinions.

Keywords: Sentiment analysis; Reddit dataset; Naive-Bayes, literature review

1 Introduction

Social media posts regarding how people are feeling can be utilized to gain a better knowledge of public health, everyday decision-making, and people's assessments of their quality of life. Evidence reveals that people are more prone to disclose their feelings online, particularly on social media platforms, during times of crisis. Sentiment analysis helps in real-time monitoring of products across public platforms. It helps in understanding customer feelings. Hackers are constantly developing new attacking tools and hacking strategies to gain malicious access to systems and attack social media networks thereby making it difficult for security administrators and organizations to develop and implement the proper policies and procedures necessary to prevent the hackers' attacks. The increase in cyber-attacks on social media platforms calls for urgent and more intelligent security measures to enhance the effectiveness of social media platforms. A more realistic idea on how well the product is doing ,since we are considering all the possible social media platforms. (eg: Reddit vs Twitter) The product owners can streamline their promotions / advertisements concentrating majorly on a specific social media platform The limitation is filtering out spam tweets from twitter and reddit because of its ambiguity

2 Related Work

Social media adversarial attack refers to a wide range of hostile operations carried out through human interactions on social media. It manipulates users' minds to make them make security mistakes or reveal important information. Attacks on social media can take several forms. To carry out the assault, a perpetrator first examines the intended victim to obtain background information such as possible avenues of entry and weak security

mechanisms. The attacker then attempts to acquire the victim's trust and give stimuli for later acts that violate security protocols, such as disclosing sensitive information or granting access to key resources.

3 Method And Evaluation

Sentiment Analysis on real-time Reddit data using Naive-bayes algorithm is the simplest and fastest technique for a large amount of data. Some of the applications for naive-bayes include sentiment analysis, text classification etc. The Naive Bayes Classifier uses the Bayes theorem to predict membership probabilities for each class, such as the probability that a particular record or data point belongs to that class. The most likely class is the one with the greatest chance of occurring.

Requirements: pandas – Python Data Analysis Library. pandas are open-source that provide high-performance, simple-to-use data structures, and data analysis tools.

Numpy – NumPy is a scientific computing fundamental package in Python. It contains among other things:

• a powerful N-dimensional array object
• sophisticated (broadcasting) functions
• tools for integrating C/C++ and Fortran code
• capabilities in linear algebra, Fourier transform, and random numbers NumPy can be used as a multi-dimensional container of generic data in addition to its apparent scientific applications.

sci-kit learn – Data mining and data analysis tools that are easy to use.

SciPy – SciPy is a Python-based ecosystem of open-source math, science, and engineering tools.

Reddit Api Keys: To fetch Reddit data

Data Preprocessing: The open available Reddit dataset on kaggle is used to train the model. For converting words into features the data preprocessing is essential. The countvectorizer here tokenizes and lower cases the text

Splitting Data: The columns are divided into dependent and independent variables first (or features and labels). The variables are then divided into train and test sets.

Vectorization: To transform each review/comment/tweet into a numerical representation

Model Generation: The initial idea was to use only Naive-Bayes algorithm for model generation. However, the results for Naive-Bayes in the first run were only 60%. In order to see if we have a better performing model the randomforestclassifier, Logistic regression and Adaboostclassifier models were tested. These models seemed to slightly overfit the model with 99% accuracy except Adaboostclassifier which was at 68%. As the data was being tested multiple times, the initial Naive-Bayes algorithm model performance improved to 70% and this was finalized. Prediction Analysis and Reddit Instance: The data from Reddit was scrapped with Reddit API key generation and getting data using PRAW and the model was able to predict 1, 0, -1 , Positive, negative and neutral consecutively

4 Data and Evaluation

Statistics: The Total number of comments/tweets in the reddit dataset are : 36,799 Negative : 8277 Neutral: 13142 Positive: 15830 The dataset used here is the open source available dataset in Kaggle to train the model. The dataset is attached as part of the code. The results are evaluated based on model accuracy scores. One interesting observation from the Real-time reddit dataset was there were many words like “suicide” , “covid” that was not retreiving the data to analyze whereas twitter was able to do that. Another observation was that different products had varying numbers for positive, neutral and negative across both platforms. The accuracy score for the model achieved was 70% for the naive bayes classification.

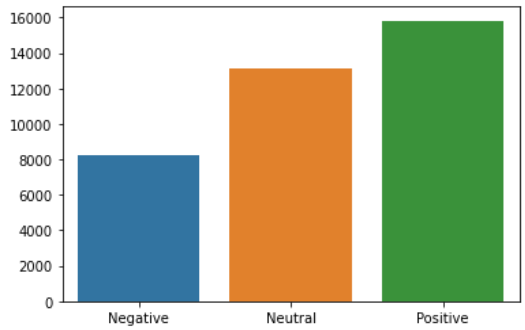

Figure 1 - Training Reddit Dataset

5 Literature Review on An Analysis of Security in Social Networks [1] - Shahana

In this research, The authors look at recent challenges to social networks. Traditional attacks are still effective in social networks, according to their findings. In the virtual world, attackers prefer to utilize social engineering to persuade consumers to visit the targeted websites. Because users are more inclined to communicate with others, attackers may carry out assaults more easily than before. The majority of attackers steal user secrets, with money being the ultimate goal in most cases. Social networking sites have also been subjected to Internet threats. People in social media tend to lower the original alarm, making it simpler for malware to propagate. In this paper, they examine contemporary threats to social networks, as well as the targets that attackers seek and the ways by which attackers carry out their operations. Users and social networking sites are separated in social networks. Then they go over the countermeasures against social network risks in depth. Finally, they present a social network security framework.

Attack and Privacy Suggested:

5.1 User countermeasures

Before joining a social networking site, users must understand the distinctions between them.

Some sites restrict access to your post to certain people, while others enable anybody to see it.

• Users must have control over the posted content.

Users can restrict access to their websites to specific groups, such as classmates, clubs, colleagues, and family

Users must have control over the posted content.

Users can restrict access to their websites to specific groups, such as classmates, clubs, colleagues, and family.

• Do not post your full name, social security number, address, phone number, bank account or credit card number, or any other user's information. Some information that potentially reveals a user's identity must be released with caution.

• Users should keep in mind that the information they upload is irreversible. Even if people delete information from websites, the outdated version of the information remains on the computers of others.

5.2 Countermeasures from social media sights

Provide various users with different functions.

Some users can only utilize a few functions by default, whereas those with more details can use more.

• When a new attack or high-risk vulnerability is discovered, users should be warned immediately. Users should also be made aware of some dubious information on the sites.

• Improve the spam and harmful link filtering.

This can prevent malicious programs from spreading and users from visiting potentially harmful websites. In addition to deploying antivirus modules, it is critical that the sites and security vendors communicate effectively.

5.3 Conclusion: A security Framework

The security of social networks is influenced by both individuals and social networking sites. Social networking services should, for the most part, provide users with adequate security protection. Users should raise their security awareness as a little part of combating the growing number of attacks.

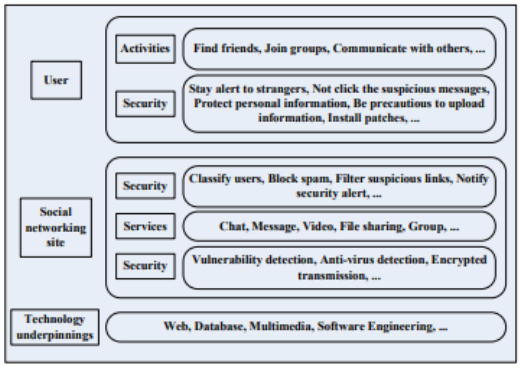

Figure 3. the security framework of social networks

Figure 2: Security framework

6 Literature Review on Adversarial Attacks and Defenses for Social Network Text Processing Applications: Techniques, Challenges and Future Research Directions [2] - Shahana

Because of the increased use of social media, numerous Machine Learning (ML) and Natural Language Processing (NLP) tools have been developed to process the massive volume of social media content and make meaningful judgments. However, adversarial attacks on these ML and NLP systems have been widely shown. Adversaries can use these flaws to conduct a variety of adversarial attacks on these algorithms in various social media word processing services. We present a complete analysis of the primary methodologies for adversarial attacks and defenses in the context of social media applications in this paper, with a special emphasis on key issues and future research prospects. Six key applications, namely (i) rumors detection, (ii) satires detection, (iii) clickbaits & spams identification, (iv) hate speech detection, (v) misinformation detection, and (vi) sentiment analysis are considered in this paper.

6.1 Learnings from the paper

Adversarial ML for social media NLP applications is a very important and growing research area as the stake is very high. For example, it allows influencing public opinions by hackers/attackers typically from state-sponsor agencies.

• Textual content in social media is very noisy, Due to such diverse characteristics, NLP in general and adversarial NLP, in particular, is more challenging. Developing defense techniques is even harder.

• Online social media applications are more prone to adversarial attacks than conventional sources of text.

• Social media applications are more susceptible to attacks and mitigation techniques are hard to integrate with social platforms as that might be seen as an attack on “freedom of speech.” So there is a delicate balance between protection against attacks, censorship, and freedom of speech.

• The widespread use of social media attracts the attackers to launch different types of adversarial attacks on these networks to fulfill their objectives.

• Social media outlets are distributed in nature, therefore, distributed attacks are easier to manifest in social media applications.

• A successful attack against the most vulnerable social media outlets might be enough to have unintended consequences (e.g., promote extreme views, etc.).

• The unstructured nature of text shared in social networks allows attackers to launch different types of adversarial attacks on the applications.

• The literature indicates that several interesting NLP applications of social networks are subject to adversarial attacks. Some of the notable applications include rumors, satires, parodies, clickbait, spam, hate speech, and misinformation detection.

• There is a need for reducing human dependency both in adversarial attack and defense

• The existing literature on adversarial NLP in general and in social media applications in particular lacks in focused benchmark datasets and quality metrics.

• Graph neural networks form a promising framework with a great potential for improving the representation of social media information.

However, this requires improving their security against adversarial attacks and mitigating their privacy leakage.

6.2 Conclusion

The authors looked at the current status of adversarial assault and defense on five key text-based social media applications: rumors, satires, clickbait & spam, hate speech, misinformation detection, and sentiment. Adversarial attacks can be used against these applications. They gave a broad review of adversarial attack and defense techniques that can be used in both text and image applications.

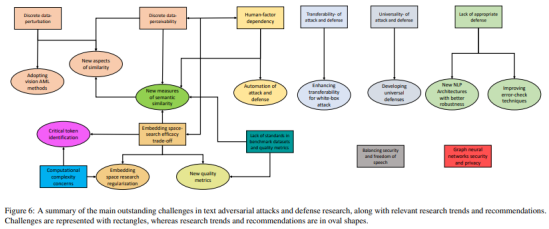

Figure 6: A summary of the main outstanding challenges in text adversarial attacks and defense research, along with relevant research trends and recommendations. Challenges are represented with rectangles, whereas research trends and recommendations are in oval shapes.

Figure 3: Attack and Defense Techniques

7 A Guide to Differential Privacy Theory in Social Network Analysis by Christine Task, Chris Clifton [3] - Sahithi

The privacy of social network data is becoming more of an issue, posing a danger to access to this vital data source. Differential privacy is a privacy model that employs noise to disguise individuals' contributions to aggregate findings and provides a very strong mathematical guarantee that individuals' presence in the data-set is hidden.Differentially-private queries inject randomized noise into query results to hide the impact of adding or removing an arbitrary individual from the data-set.

7.1 Learnings from the paper

They created a new standard for differential privacy called out-link privacy that provides strong privacy guarantees with the introduction of very small noise. They provided straightforward applications to differential privacy. They analyzed the feasibility of the pre-existing social network techniques under the differential privacy standards. There is going to be an issue with the trade off between the potential privacy and utility.There might be stronger privacy with more noise being generated ,but if there is more noise then the dataset would not function properly. These techniques satisfy a weaker definition of differential privacy, and in some cases computing how much noise is required to privatize a given DI may be infeasible.

We think that by making this guide to differentially private social network analysis available, as well as new, powerful approaches for privatizing social-network data, we may encourage the practical application of these standards to social-network data. We intend to investigate the applicability of outlink privacy to other social-network analytic tasks in the future and offer results. These methods have been tested on real-world network data.

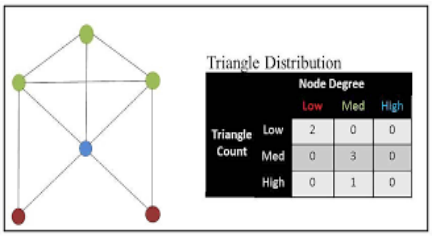

Figure 4: Triangle distribution

8 Literature Review on Evasion attacks against machine learning at test time Battista Biggio1 , Igino Corona1 , Davide Maiorca1 , Blaine Nelson2 , Nedim Srndi´c ˇ 3 , Pavel Laskov3 ,Giorgio Giacinto1 , and Fabio Roli1 [4] - Sahithi

This paper introduces the problem and the working of creating evasion attacks that use the adversarial examples and are created with the level of confidence which talk about the minimum distance misclassifications and the substitute models. These

are the two basic concepts that are used and also reused for these attack's development against the deep networks.

8.1 Learnings from the paper

The attack strategy is really good because it is based on the nonlinear optimization problem. The techniques used are gradient descent or quadratic techniques which include Newton's methods.

The proposed method is demonstrated using two examples such as the toy example on handwritten digits and malware detection in pdf files. The authors proposed that the attacks raised questions about how the detection of the malware in pdfs can be feasible and have solved its versioning mechanism and implemented legitimate Open action code. The enclosure of the legitimate samples might have the issue of increasing the risks to mimic the legitimate class and this can't always be possible. There might be problems where the attack can take place at the distribution or can add the generated attack samples to the training set. The paper could venture on how to distribute the model by adding the generated samples of the attack to the training set. By improving the security the paper can probably be balanced with the
higher FP rate

9 Results and Insights

Naive Bayes model has an accuracy score of 70% and performs well on reddit data. We retrieved 500 real-time reddit posts for alexa and siri subreddits and classified them as positive, neutral and negative. Through our papers we also found that in social media we can perform character level attacks, word level attacks, sentence level attacks, multilevel text adversarial attacks. Defense includes adversarial training, defensive distillation, recovery of perturbation etc.

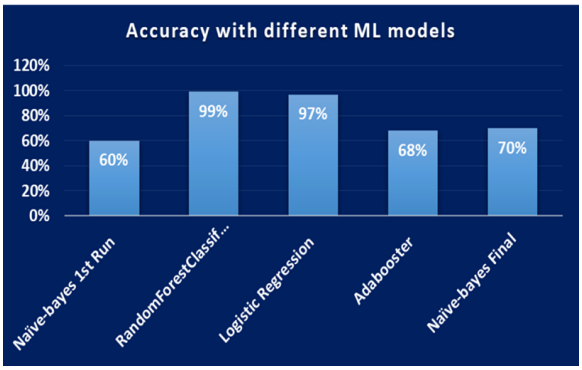

Figure 5: ML model accuracy

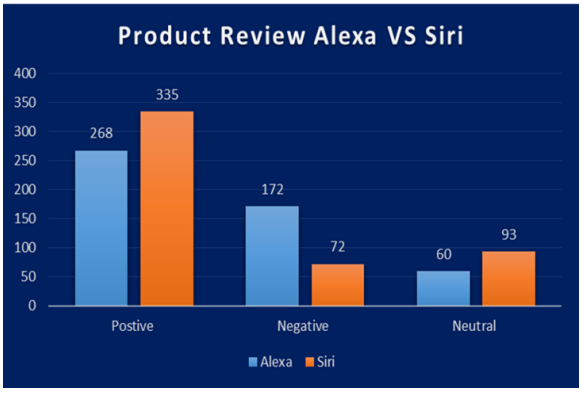

Figure 6: Alexa Vs Siri sentiment analysis

10 Future Improvements

The model can be created to work on multiple social media platforms such as twitter, facebook and instagram. We can introduce differential privacy in the form of a training bot ,what this essentially means is that during the introduction of differential privacy there is an introduction to noise , this noise can be generated using bots on the reddit dataset since it is based on which product is better to advertise on which platform and need user's personal details. We observe issues with the security and These bots can create fake credit card information or phone numbers and introduce the attacks which we can later use the defense for and then apply into the real life situation. We also need to create a more robust defense mechanism for these social media platforms with large datasets.

11 Task Contribution

Name	Task Performed
Shahana	Model Creation + Literature review 1 and 2
Sahithi	Model Creation + Literature review 3 and 4

Table 1: Work contribution

12 References

[1] An Analysis of Security in Social Networks Weimin Luo1 , Jingbo Liu1 , Jing Liu2

[2] Adversarial Attacks and Defenses for Social Network Text Processing Applications: Techniques, Challenges and Future Research Directions Izzat Alsmadia , Kashif Ahmadb , Mahmoud Nazzalc , Firoj Alamd , Ala Al-Fuqahab , Abdallah Khreishahc , Abdulelah Algosaibie

[3] A Guide to Differential Privacy Theory in Social Network Analysis by Christine Task, Chris Clifton

[4] Evasion attacks against machine learning at test time Battista Biggio1 , Igino Corona1 , Davide Maiorca1 , Blaine Nelson2 , Nedim Srndi´c ˇ 3 , Pavel Laskov3 ,Giorgio Giacinto1 , and Fabio Roli1

[5] Adversarial Attack on Sentiment Classification Alicia Yi-Ting Tsai, Tobey Yang

[6] Privacy and security in online social networks: A survey Imrul Kayes a,* , Adriana Iamnitchi b