УДК 004.49

DETECTION OF MASKED BAD WORDS IN SOCIAL MEDIA CONTENT

Dube Elton M.

Social Media, is it a blessing or a curse? Many people will have different views on this matter, but one can never ignore the fact that it has a very big influence in people's lives in modern times. With the increase in popularity and availability of different social media platforms, more and more people are finding it easier and easier to communicate with each other all over the world. That is just one of the highlights of social media.

Now let us look at the downside of social media. While many people use it for simple communication with friends and family and keeping up with the latest trends and news, others on the other hand are using it for all the wrong reasons including bullying and circulating false information.

All of this has led to the development and improvements of things such as the detection of abusive language, hate speech, cyberbullying, and trolling amongst others. Social Media Sites are being tasked to continuously improve their cybersecurity measures to protect their users from cyberbullying.

With the world ever evolving and humans becomes more and more sophisticated and intelligent, this has led to cyberbullies adapting to these restrictions put in place by social media sites and now masking bad words, profanity and hate speech, and this has made it hard for some models to detect some bad words and profanity.

With the help of machine learning we have come up with ways to detect masked bad words in social media content and in this report, we will look at some models that are being used and how we can further improve them in the future.

Work to be done:

We will first need to source out a vocabulary of bad words that we can use as part of our available datasets**.**

After identifying a vocabulary of bad words, we will now focus on analyzing individual isolated words from a comment with the help of embedding.

We will use the Levenshtein Edit Distance to check if there is a direct match of the isolated word with the bad word list. If there is a direct match, then the algorithm will break.

If there is no direct match, then it should calculate Levenshtein Distance and form a Candidate list.

A major part of the research work is going to be finding a new procedure for the Levenshtein Distance.

This distance will be calculated considering visual similarity of a pair of symbols e.g., the pair Y and B – visual similarity = 0, but for the pair B and 8 – visual similarity = 0.85 and for the pair S and 5 – visual similarity = 0.7.

For this we will need to find a confusion matrix for these symbols and then incorporate it into the Levenshtein Distance.

So, a brief summary of the task is that we will find a confusion matrix that we can use to incorporate into the Levenshtein Distance so that we can be able to calculate the distance considering visual similarity of the isolated word and the bad word vocabulary list.

References

1. Sara Owsley Sood, Judd Antin, Elizabeth F Churchill. Using Crowd sourcing to improve profanity Detection. *AAAI Technical Report SS-12–06 Wisdom of the Crowd*

2. Aouragh Si Lhoussain, Gueddah Hicham, Yousfi Abdellah. Adapting the Levenshtein Distance to contextual spelling correction. *International Journal of Computer Science and Applications.* March 2015.

3. Bouma H. Visual recognition of isolated lower-case letters, Vision Research, 1971, 11, 459–474

4.Sandip Modha, Thomas Mandl, Prasenjit Majumder, Daksh Patel, DA-IICT,Gandhinagar, India sjmodha@gmail.com , Overview of the HASOC track at FIRE 2019: *Hate Speech and Offensive Content Identification in Indo-European Languages*, FIRE 2019, 12–15 December 2019, Kolkata, India.

5. Shervin Malmasi, Marcos Zampieri. *Detecting Hate Speech in Social Media*, 26 Dec 2017, doi:arXiv:1712.06427v2 [cs.CL].

6. Arup Baruah, Ferdous Ahmed Barbhuiya, Kuntal Dey: IITG-ADBU at HASOC 2019: *Automated Hate Speech and Offensive Content Detection in English and Code Mixed Hindi Text*, 2019, pp. 12–15, doi: http://ceur-ws.org/Vol-2517/T3-7.pdf

7. Vijayasaradhi Indurthi1,3, Bakhtiyar Syed1, Manish Shrivastava1 Manish Gupta1,2, Vasudeva Varma11 IIIT Hyderabad, 2 Microsoft, 3 Teradata Fermi at SemEval-2019 Task 6: *Identifying and Categorizing Offensive Language in Social Media using Sentence Embeddings*

8. John Pavlopoulos, Ion Androutsopoulos, Nithum Thain, Lucas Dixon ConvAI at SemEval-2019 Task 6: *Offensive Language Identification and Categorization with Perspective and BERT*

УДК 004.056.55

АТАКИ НА СТЕГАНОСИСТЕМИ. КРИПТОГРАФІЧНІ АТАКИ

П. В. Римар, В. В. Крохмалюк

Захист інформації від несанкціонованого доступу вирішуються в усі часи історії людства. Вже в стародавньому світі виділилося два основні напрямки вирішення цієї задачі, що існують і до сьогоднішнього дня: криптографія і стеганографія. Метою криптографії є приховання вмісту повідомлень за рахунок шифрування. На відміну від цього, при стеганографії приховується сам факт існування таємного повідомлення.

Серед можливих методів (заходів) захисту конфіденційної інформації найпоширенішим на сьогодні є метод криптографічного захисту, під яким розуміється приховання змісту повідомлення за рахунок його шифрування (кодування) за певним алгоритмом, що має на меті зробити повідомлення незрозумілим для непосвячених в цей алгоритм або у зміст ключа, який використовувався при шифруванні. Але зазначений метод захисту є неефективним щонайменше з двох причин.

По-перше, зашифрована за допомогою більш-менш стійкої криптосистеми інформація є недоступною (протягом часу, що визначається стійкістю криптосистеми) для ознайомлення без знання алгоритму і ключа.

По-друге, слід звернути увагу на те, що криптографічний захист захищає лише зміст конфіденційної інформації. У цьому випадку проблема інформаційної безпеки повертається до стійкості криптографічного коду.

На противагу вище зазначеному, стеганографічний захист забезпечує приховання самого факту існування конфіденційних відомостей при їх передачі, зберіганні чи обробці. Під приховуванням факту існування розуміється не тільки унеможливлення виявлення в перехопленому повідомленні наявності іншого (прихованого) повідомлення, але й взагалі зробити неможливим викликання на цей рахунок будь-яких підозр. Загальною рисою стеганографічних методів є те, що приховуване повідомлення вбудовується в деякий не приваблюючий увагу об’єкт (контейнер), який згодом відкрито транспортується (пересилається) адресату.

Завданням пропонованої роботи є розробка програмного комплексу для демонстрації принципів, закладених в основу поширених на сьогодні методів стеганографічного приховування з можливістю обчислення основних показників спотворення контейнера при вбудовуванні до нього приховуваних даних.

Дана задача вирішується попереднім опрацюванням наступних питань:

– розгляд особливостей побудови стеганографічних систем та основних типів атак на зазначені системи;