Be Safe & Play Fair

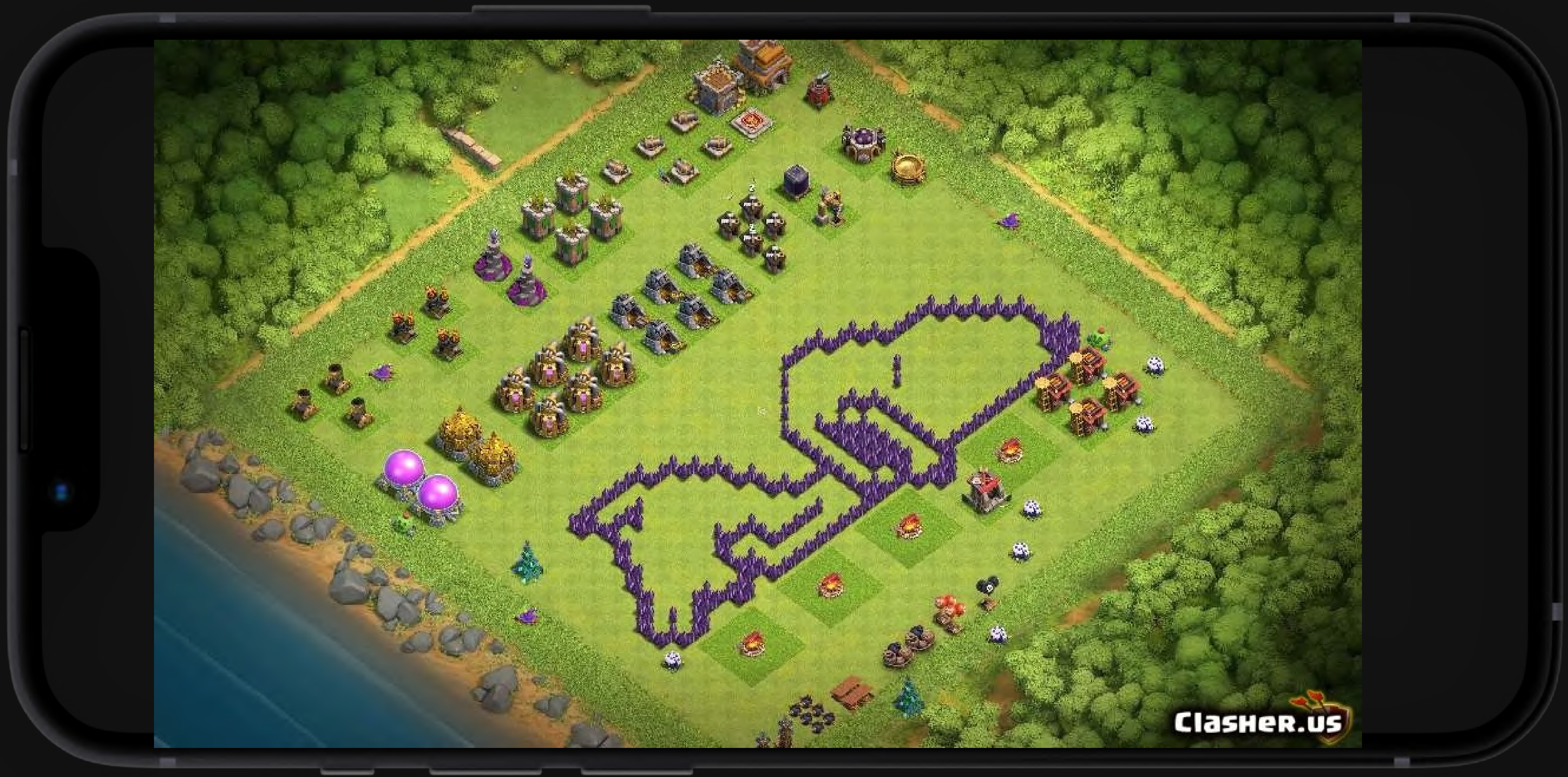
INSPECT

SUPERCELL

Let’s see what El Primo is doing..

SUP
ERC
ELL

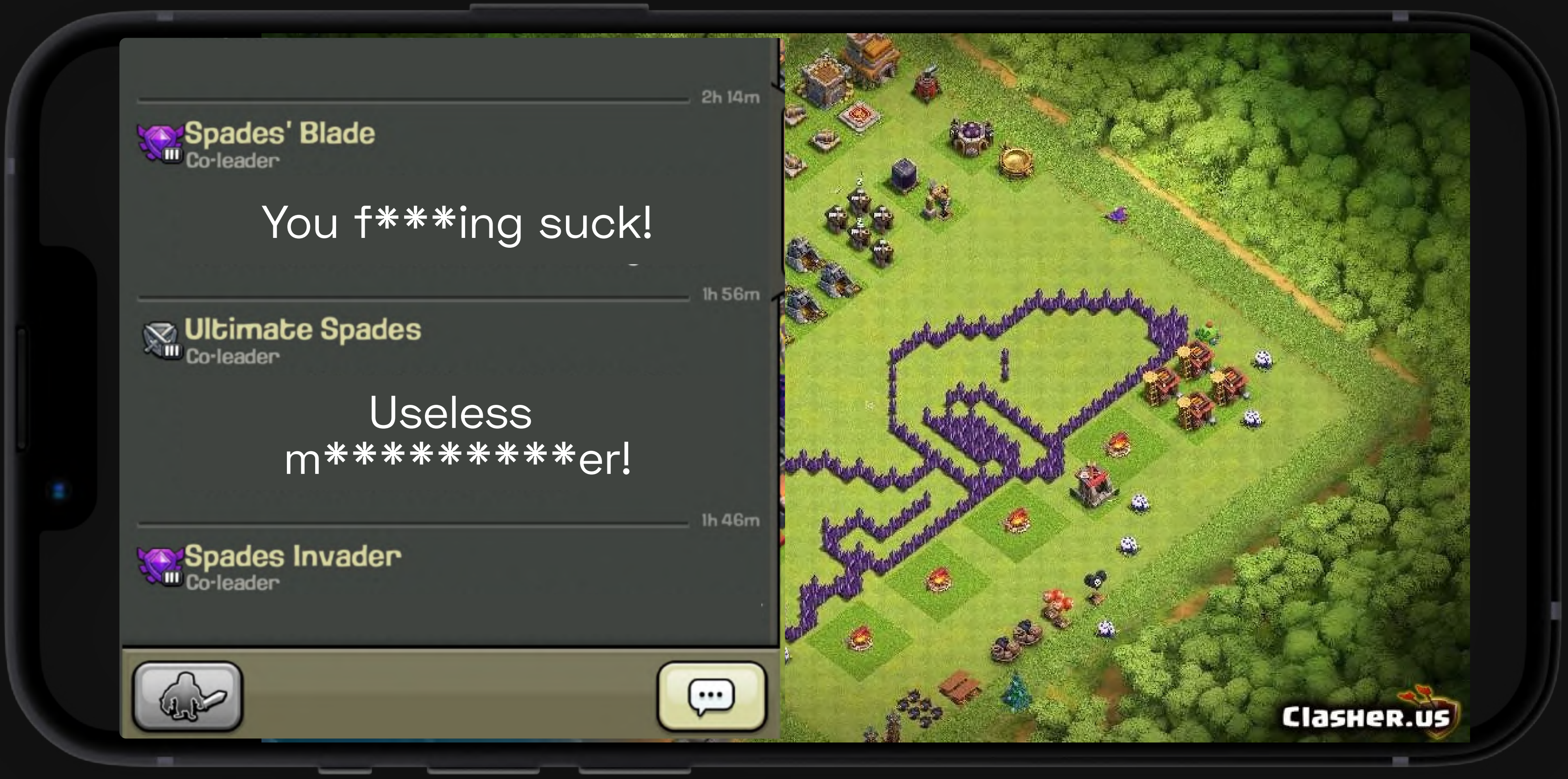
Let’s see what El Primo is doing..



INSPECT

Try Pitch

Let’s see what El Primo is doing..



2h 14m

Spades' Blade
Co-leader

You f***ing suck!

1h 56m

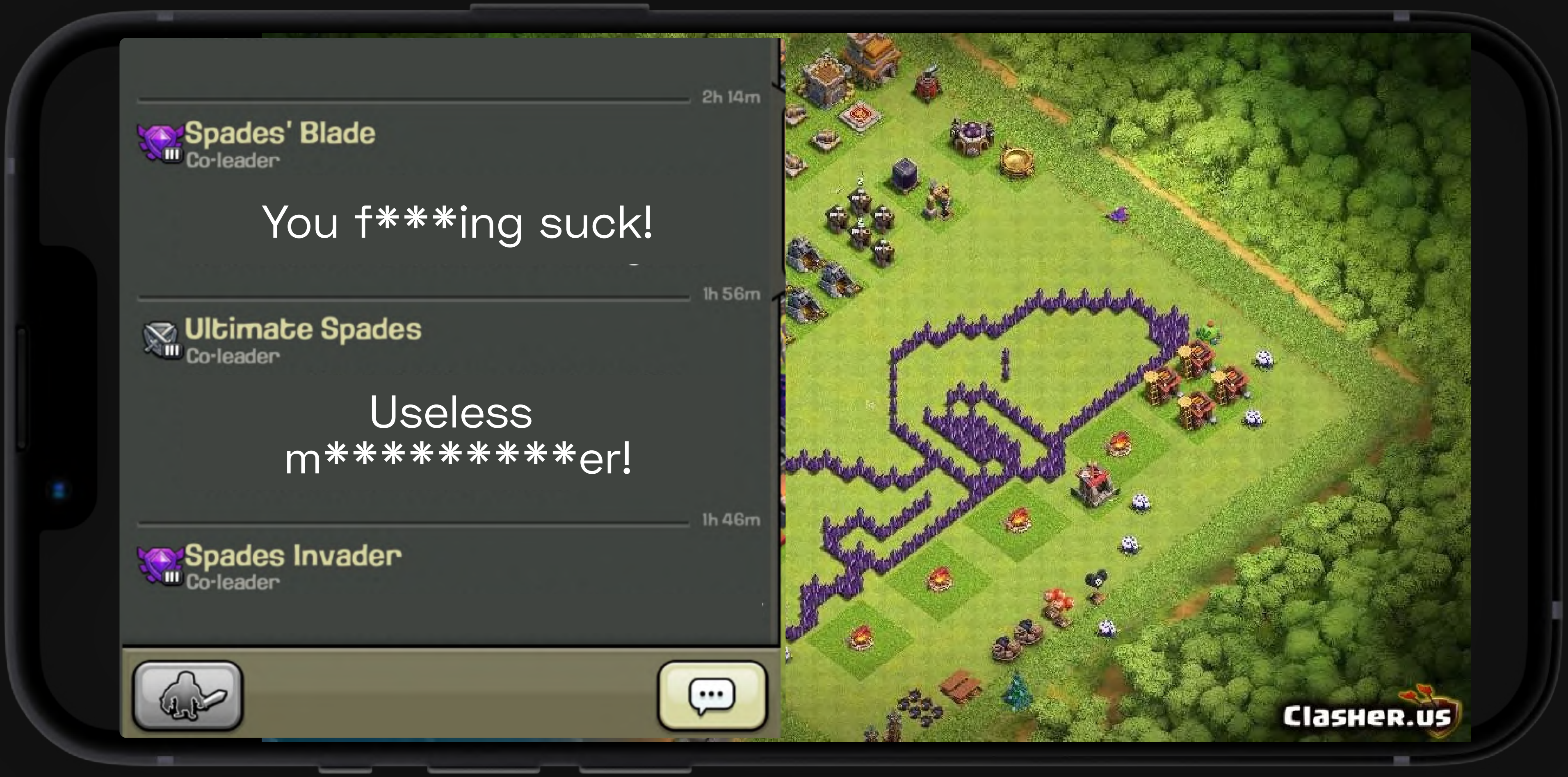
Ultimate Spades
Co-leader

Useless m*********er!

1h 46m

Spades Invader
Co-leader

Let's see what El Primo is doing..


2h 14m

Spades' Blade
Co-leader

You f***ing suck!

1h 56m

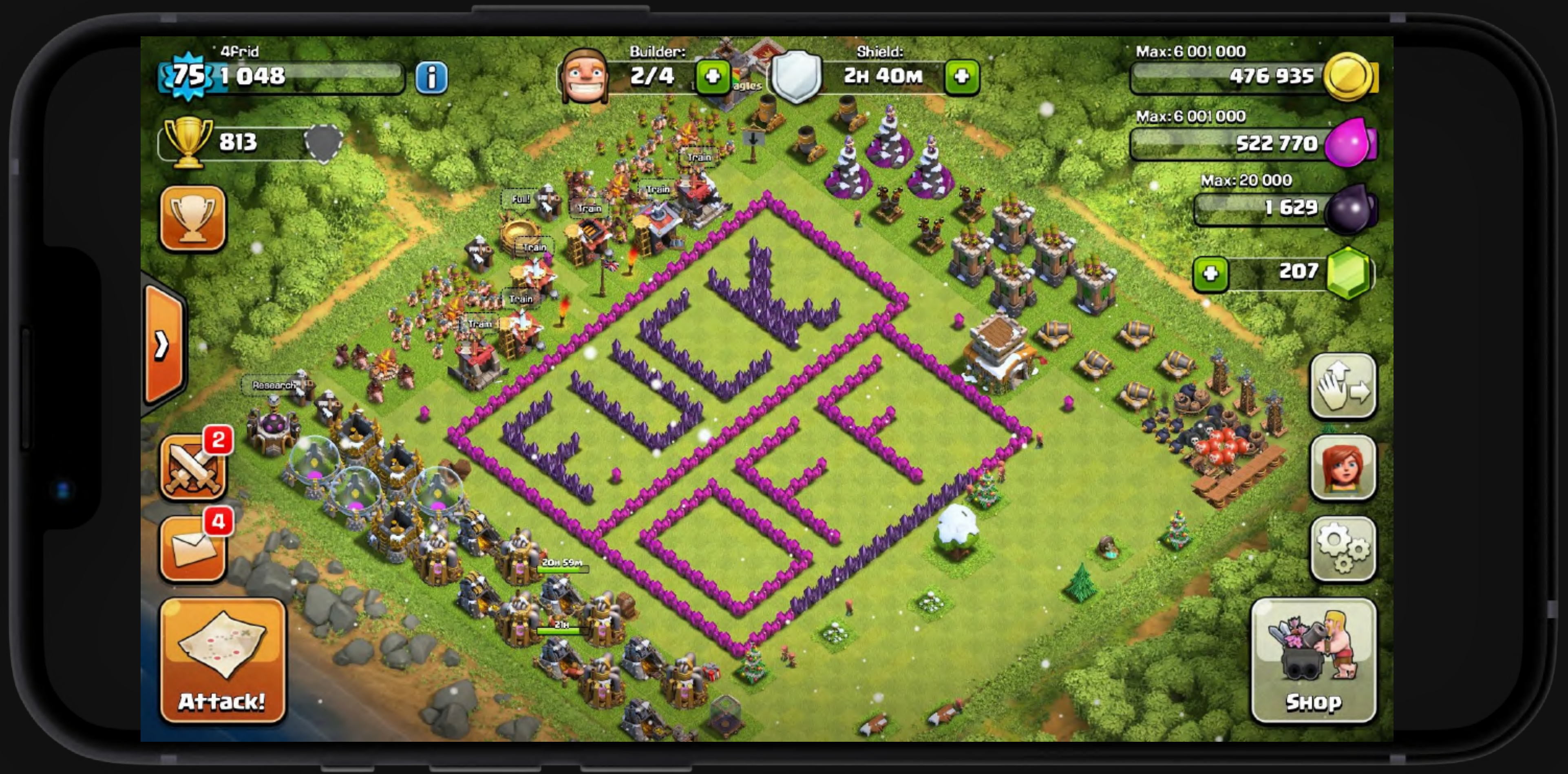
Ultimate Spades
Co-leader

Useless m**********er!

1h 46m

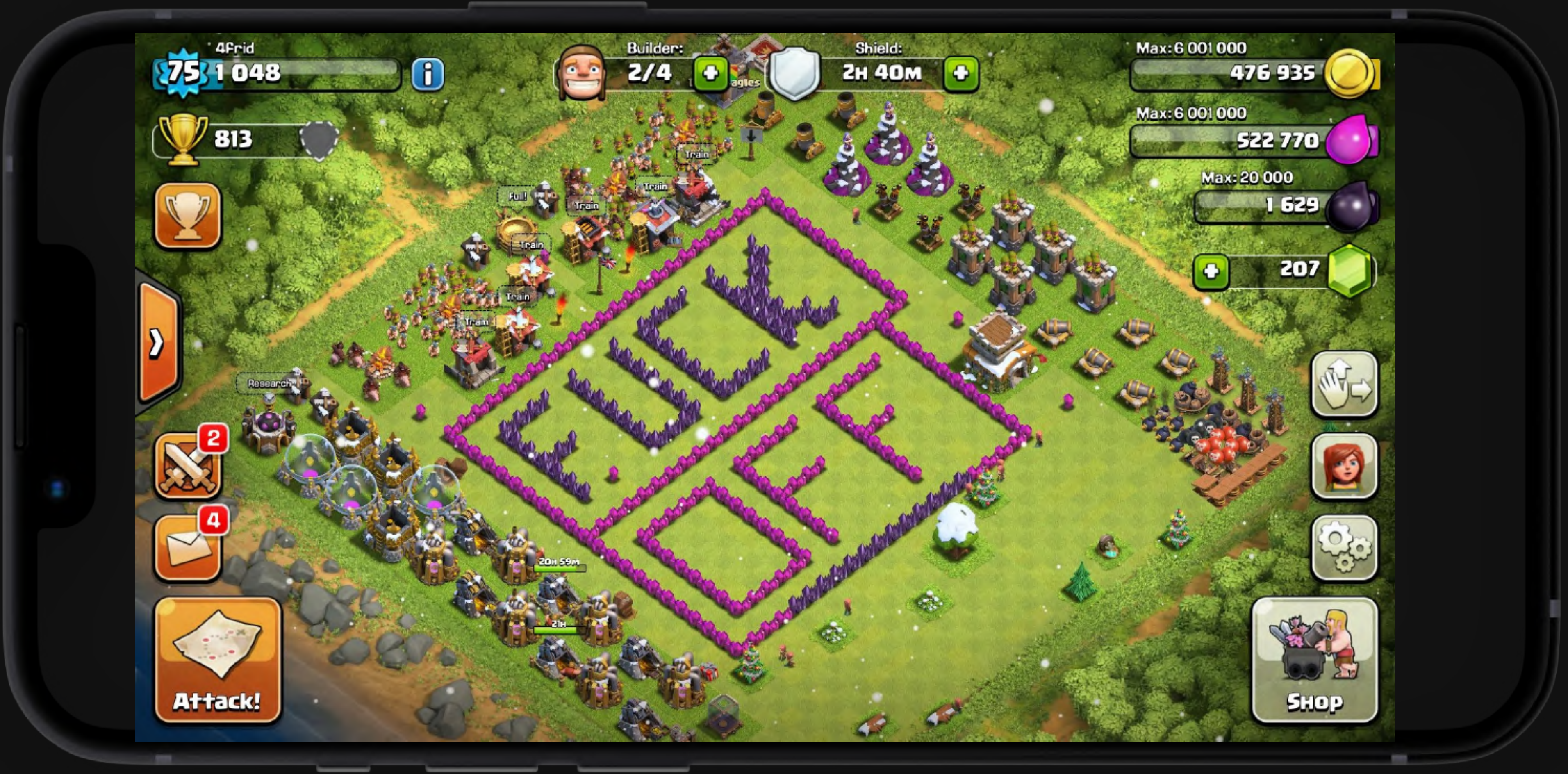
Spades Invader
Co-leader

Let’s see what El Primo is doing..

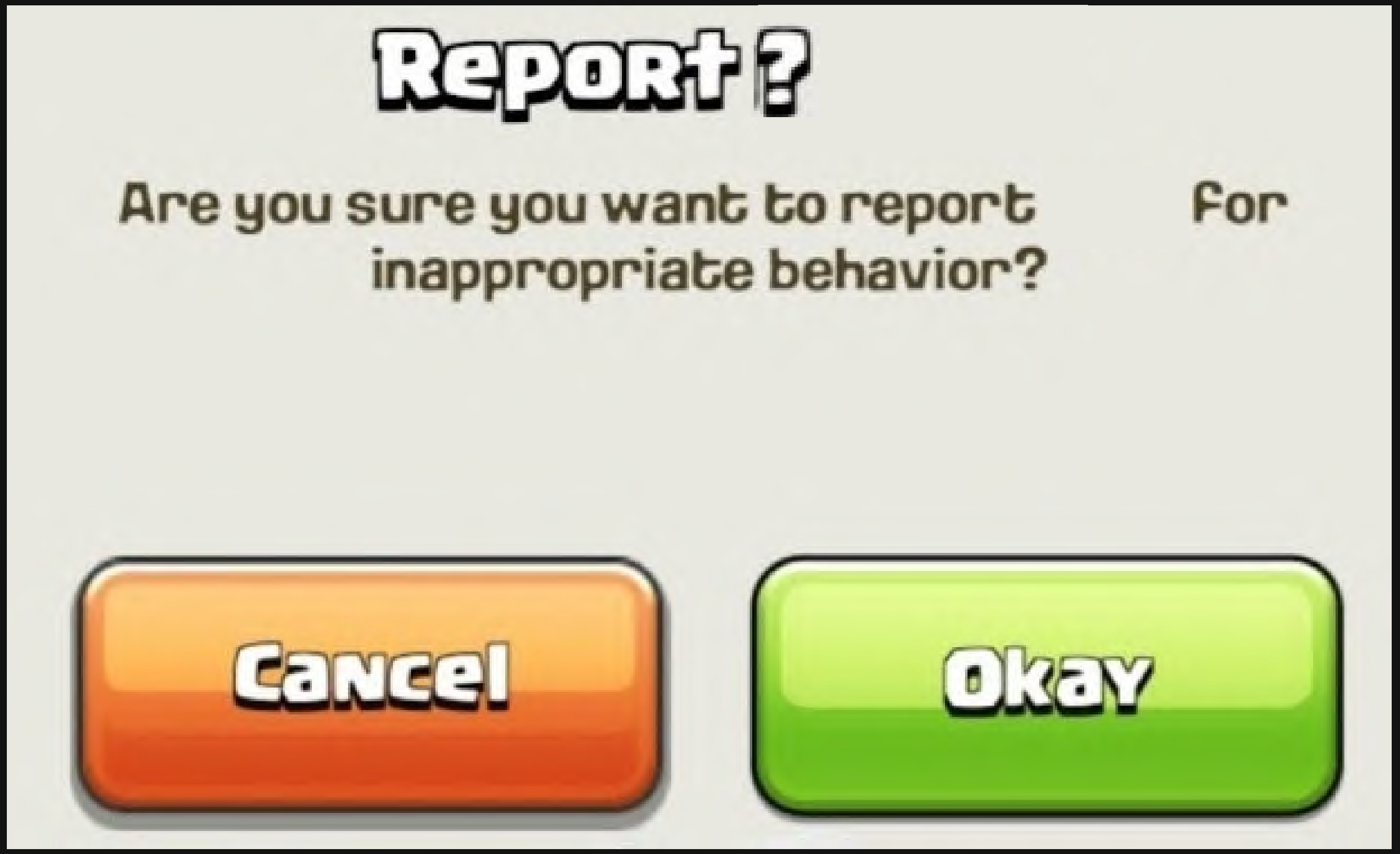


INSPECT

Let’s see what El Primo is doing..



Should el Primo report or not?



Report ?

Are you sure you want to report for inappropriate behavior?

Current problems

1. AI cannot solve the problem yet, chat in groups
2. Users not reporting ➤ **Increase reporting fluidity**
3. Abusers ➤ **Reduce propensity to abuse**
4. Moderators cost ➤ **Increase efficiency & accuracy**

INSPECT

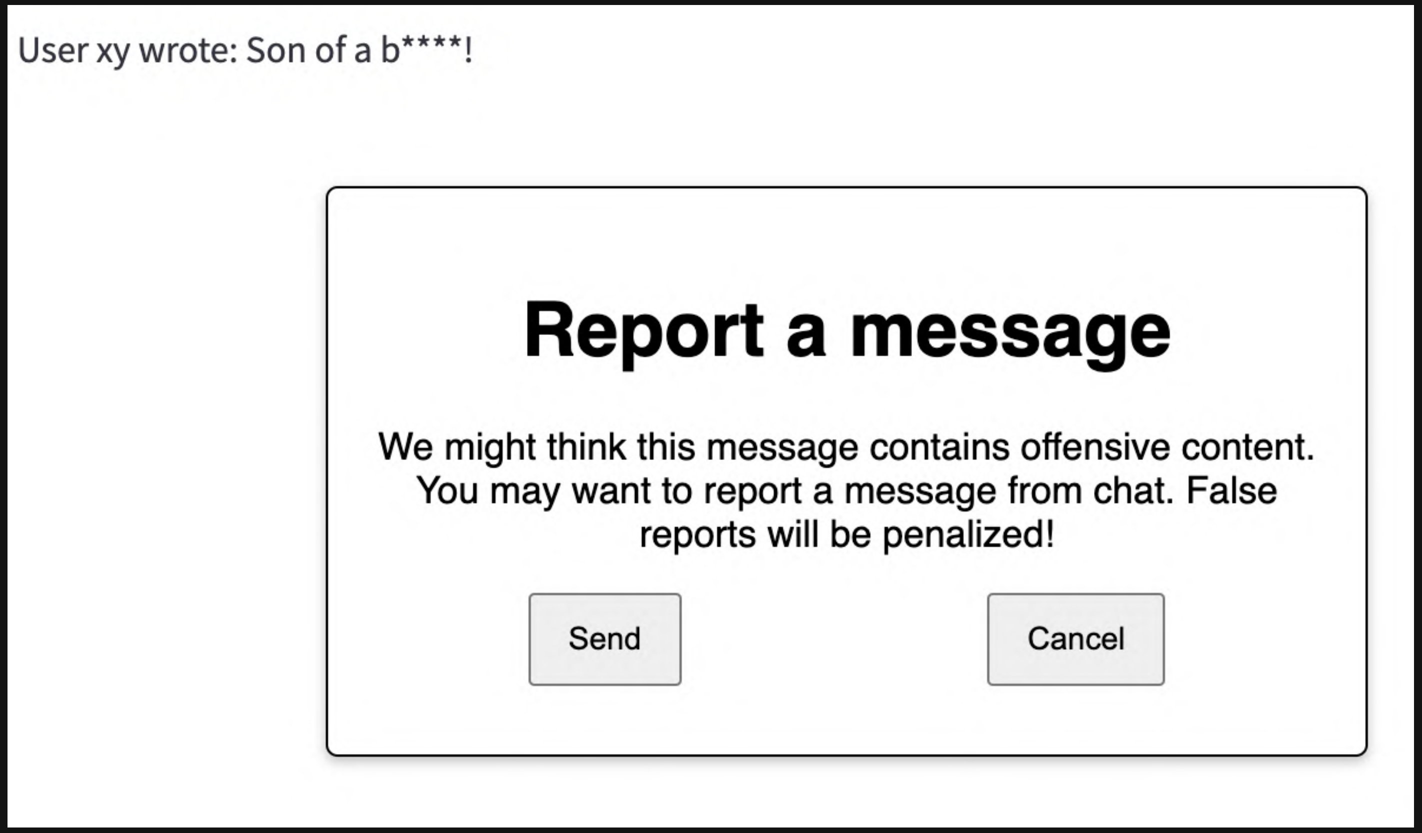
The three-pillars strategy

1. Prompt chat reporting
2. Base-layout filtering
3. Rewarding system

Two-sided Prompted chat reporting

- Offensive Speech Detection with RoBERTa (pretrained)
- Accuracy of 81% across 70k messages from chats

Victim view

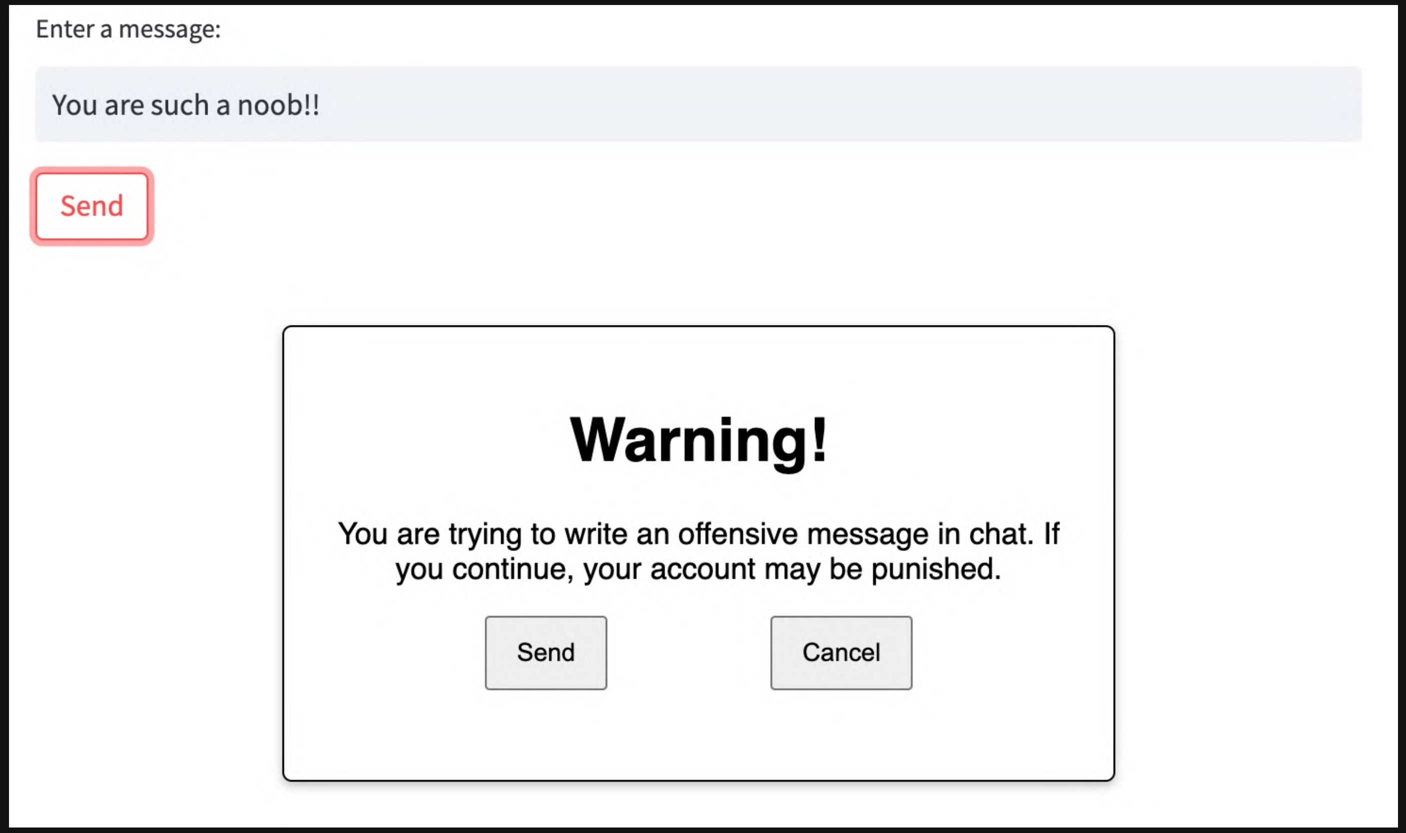


User xy wrote: Son of a b****!

Report a message

We might think this message contains offensive content. You may want to report a message from chat. False reports will be penalized!

Bullies view



Enter a message:

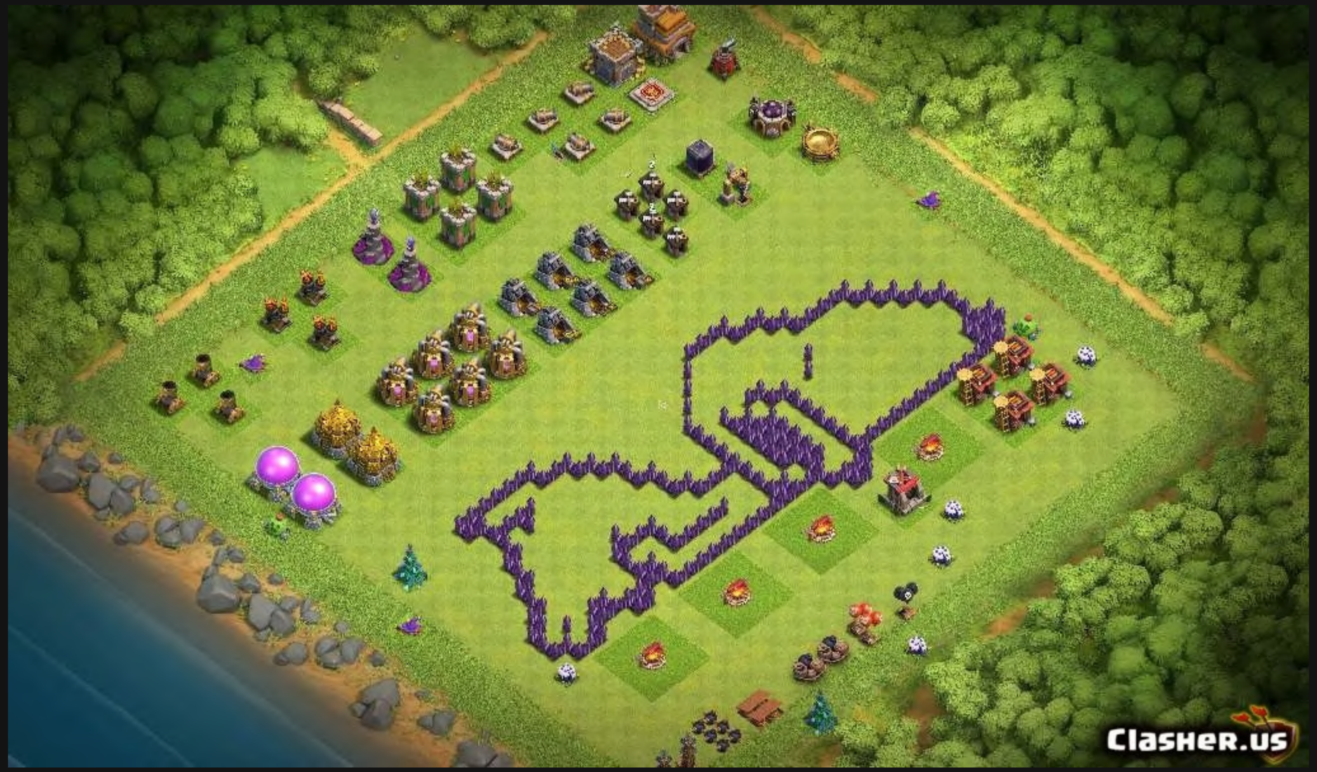
You are such a noob!!

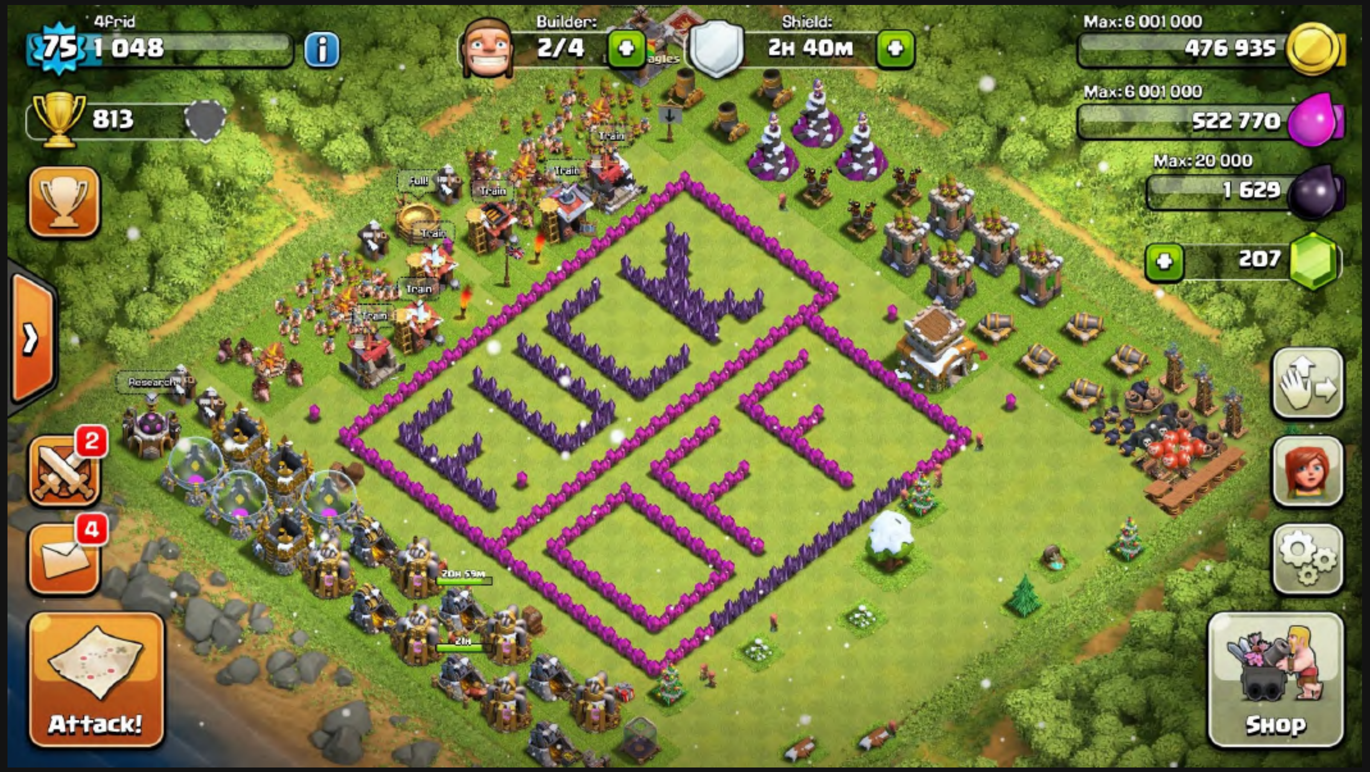
Warning!

You are trying to write an offensive message in chat. If you continue, your account may be punished.

INSPECT

Vision Transformer for offensive bases detection

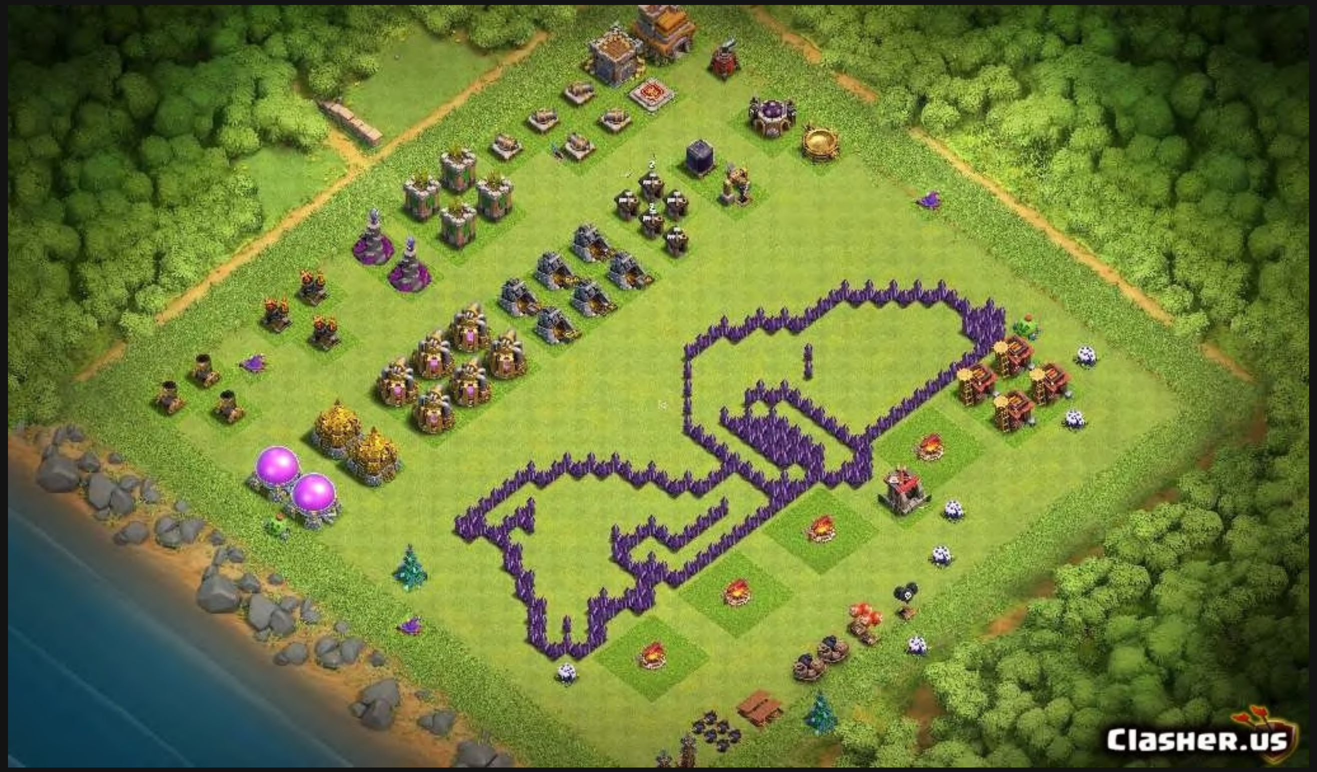


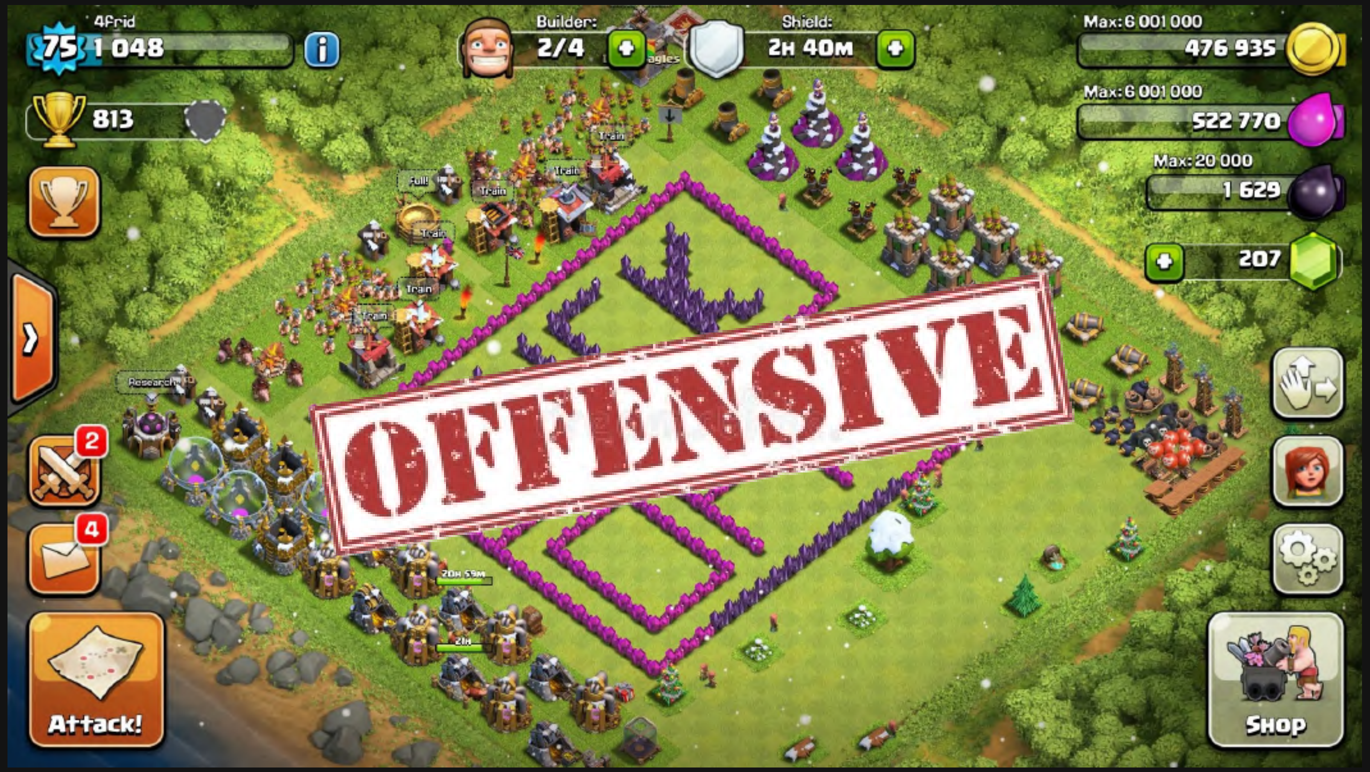


- Self-collected dataset (approx. 250 images)
- Vision Transformer pretrained on ImageNet
- Fine-tuning of 2 epochs and freezing last layers

SUP
ERC
ELL

Vision Transformer for offensive bases detection





- Self-collected dataset (approx. 250 images)
- Vision Transformer pretrained on ImageNet
- Fine-tuning of 2 epochs and freezing last layers

INSPECT

Reputation system - key points

1. Play longer and play fair - linear improvement of reputation
2. Offensive behaviours decrease the score
3. Successful reporting - increases score, false reporting - decreases the score
4. Most reputable players help with community interaction and gain rewards, helps moderators.

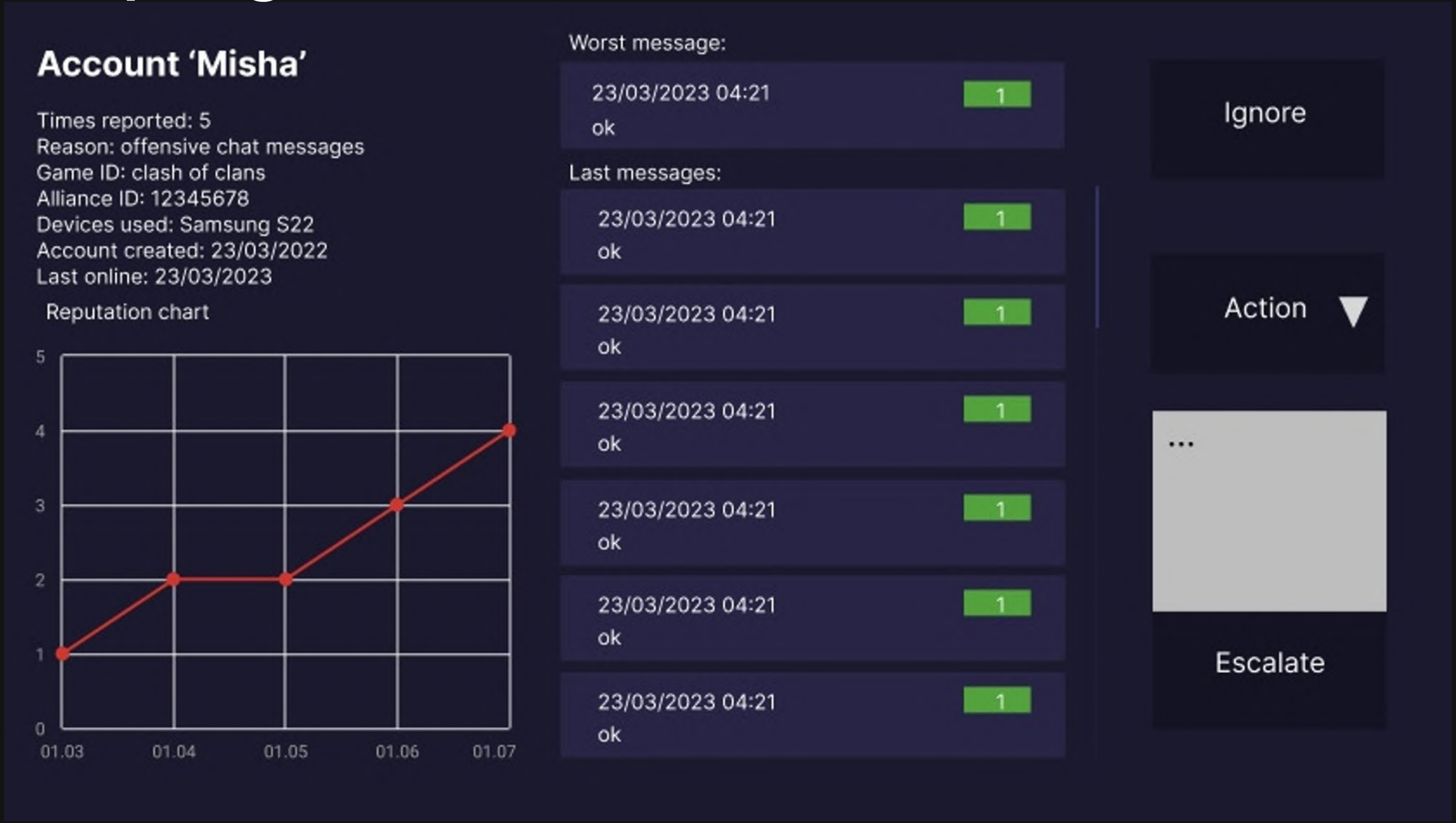
Refer to our documentation for a complete overview of the reward system we envision

INSPECT

Try Pitch

SUP
ERC
ELL

Helping Moderators



Account 'Misha'

Times reported: 5
Reason: offensive chat messages
Game ID: clash of clans
Alliance ID: 12345678
Devices used: Samsung S22
Account created: 23/03/2022
Last online: 23/03/2023

Reputation chart

Worst message:

23/03/2023 04:21 1
ok

Last messages:

23/03/2023 04:21 1
ok

23/03/2023 04:21 1
ok

23/03/2023 04:21 1
ok

23/03/2023 04:21 1
ok

23/03/2023 04:21 1
ok

23/03/2023 04:21 1
ok

INSPECT

SUPERCELL

Let’s fight offensive cyber behaviour together!

Extend the current offensive-cyber-behaviour detection system to the base-layout and integrate a reward system to nudge good behaviour instead of banning bad actions

INSPECT