Toxicity in twitch

Lluís Hernández
Sergi Olives
Eloi Yerpes

#SomTalaiòticsin'Eloi
Universitat Pompeu Fabra
Novembre 2019

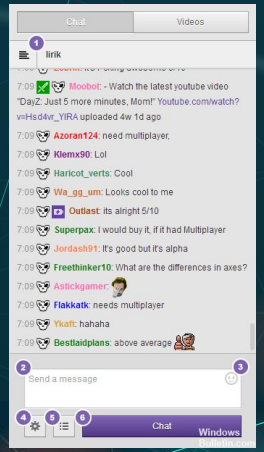
CONTENTS

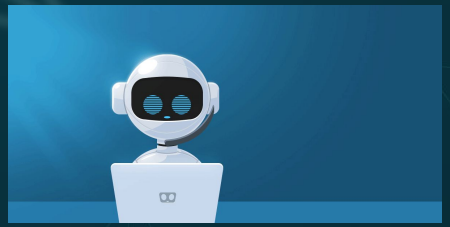
MAIN GOAL

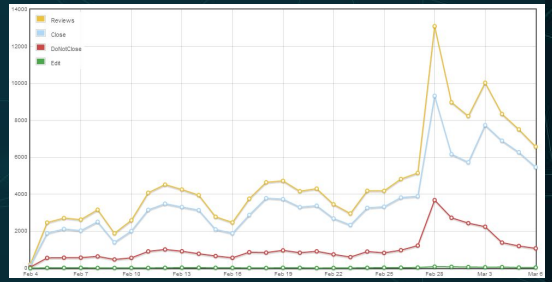
Main Goal: **REDUCE CHAT TOXICITY IN TWITCH**

How to do it:

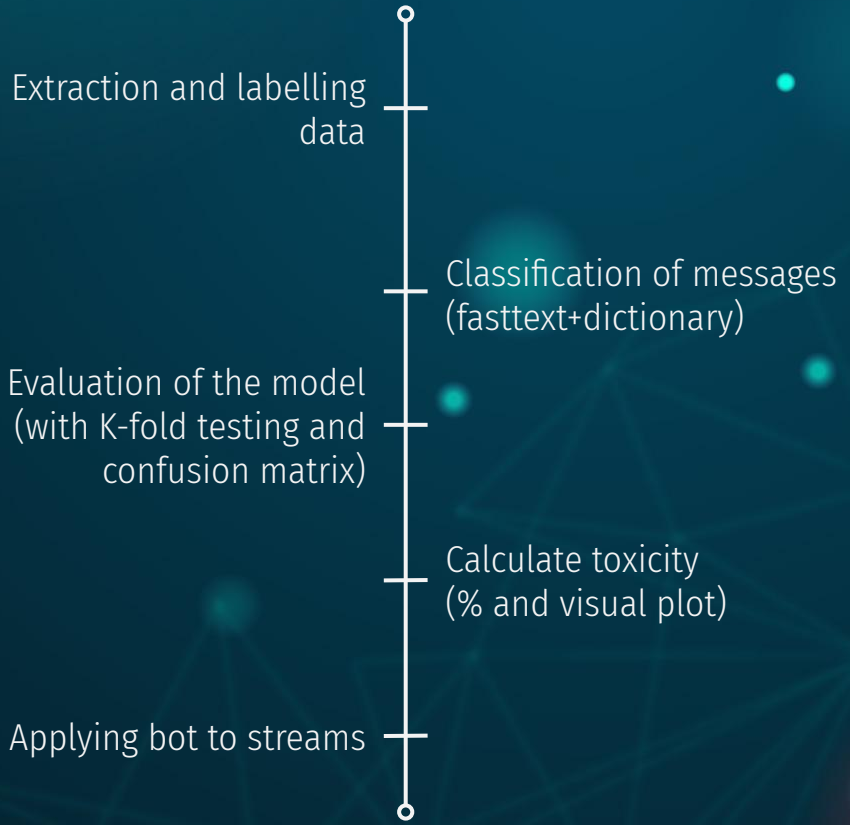
- Analyze chat toxicity in different streams
- Create an application in Twitch to reduce toxicity
- Display information about chat toxicity in a visual manner for streamers







EVOLUTION of the process



DATA EXTRACTION & LABELLING

Data collection:

- Twitch bot reading chat

Classify each message manually in:

- Toxic
- Not Toxic

Criterion?

Corpus of 975 messages:

➔ 70% Not toxic
➔ 30% Toxic

NORMALIZATION OF MESSAGES

Each message used in the training model is normalized.

Each message being predicted from a live twitch chat is also normalized.

Normalization process:

- Lowercase
- Removing signs as [!”#$%&’()*+,-./:;<=>?@[\]^_`{|}~]:
- Stripping the string (remove whitespaces, tabs and line jumps)

U can flip my Shit :) → u can flip my shit

FASTTEXT

FastText: library for text classification and representation

Text -> continuous vectors

Python Language for the whole project

Create a fasttext supervised model:

- We train it with our train set of 975 **NORMALIZED** labeled msgs.
- We use 25 epochs, 2 wordNgrams and a lose rate of 1.0.

```
__label__toxic i hope you die
__label__toxic fuck
__label__toxic show bob
__label__toxic stoopid
__label__toxic can u just suck it
__label__toxic you're a furry pog
__label__toxic n*gg
__label__toxic shroudw fuck ya
__label__toxic trash game
__label__toxic u suck
__label__toxic i have aids
__label__toxic fuckin dog
__label__toxic u dumb
```

```
__label__notoxic heyyyy
__label__notoxic lets go boys
__label__notoxic ben trolling
__label__notoxic i believe in you
__label__notoxic it's over
__label__notoxic ofc he hits it
__label__notoxic bennnyyy
__label__notoxic bruh crit is lagging
__label__notoxic get your shirt back off
__label__notoxic motivate him
```

EVALUATION OF THE FASTTEXT MODEL

We iterated the training model 3 times:

• 483 messages without normalization	60% NToxics 40% Toxics	→ Accuracy: 68%
• 483 messages with normalization	60% NToxics 40% Toxics	→ Accuracy: 71%
• 975 messages with normalization	70% NToxics 30% Toxics	→ Accuracy: 83%

Stratified K-Fold testing with K = 10 with confusion matrix

COMPARISON OF MODELS

483 Messages Accuracy: 71%		Actual	
		No Toxic	Toxic
Prediction	No Toxic	234	41
	Toxic	99	109

No Toxic
Precision: 85% Recall: 70% F1: 77%

Toxic
Precision: 52% **Recall: 73%** F1: 61%

975 Messages Accuracy: 83%		Actual	
		No Toxic	Toxic
Prediction	No Toxic	663	55
	Toxic	110	148

No Toxic
Precision: 91% **Recall: 86%** **F1: 89%**

Toxic
Precision: 58% Recall: 71% **F1: 64%**

ANALYZING ERRORS IN THE MODEL

There are more toxic messages marked incorrectly as non-toxic than non-toxic messages marked as toxic.

Curious examples:

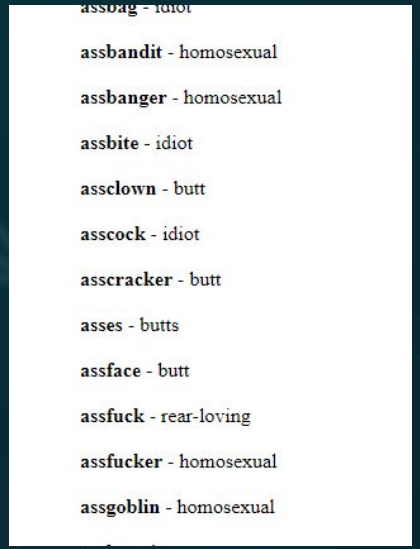
- Messages containing "**ur**" get always **toxic** because of instances as **"ur gay", "ur bad"**...
- Messages marked incorrectly as toxic can be caused by **different meanings in different contexts**
- Messages marked incorrectly as non-toxic are usually because of misspelling and orthographic errors (by error or consciously)

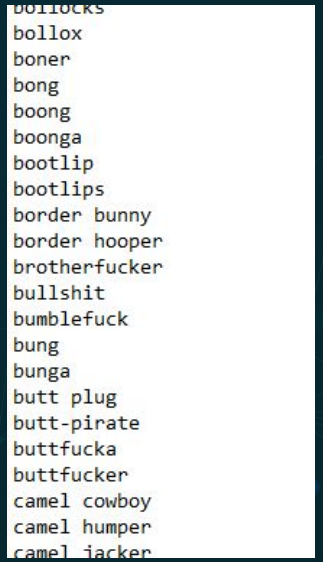
Fasttext also informs about the **certainty of the prediction** with a % and there are **not a noticeable difference between correct predictions and incorrect ones.**

DICTIONARY

Dictionary containing more than 200 words:

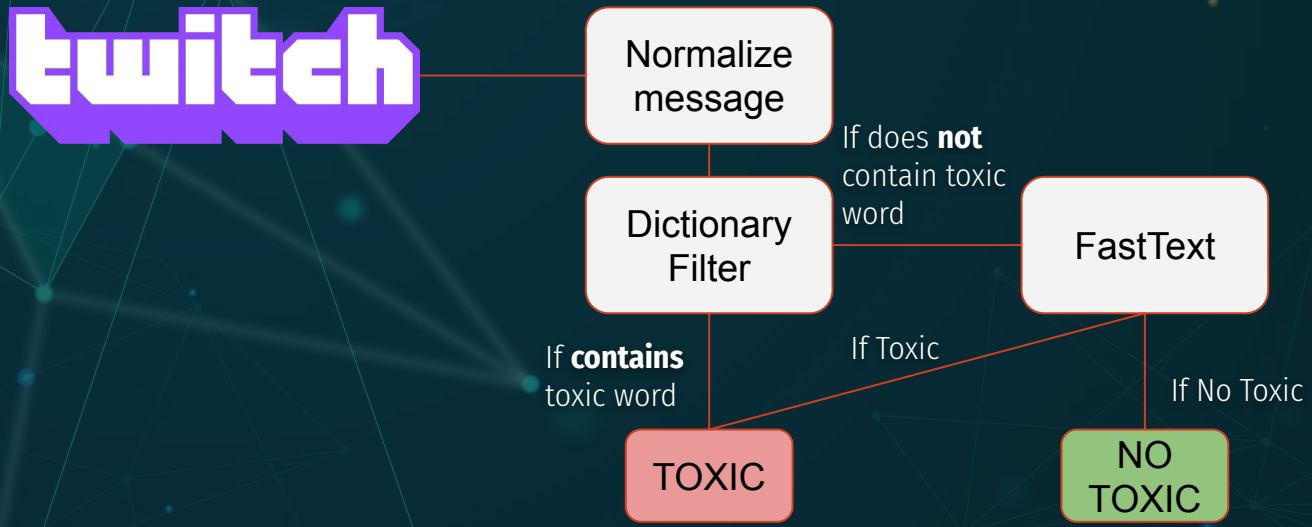
- **www.noswearing.com/dictionary**
- **https://hatebase.org/**





TOXICITY DETECTION PIPELINE

Unify in the same script both dictionaries and Fasttext to find toxic messages:



Adding this extra dictionary filter helps to get a faster detection of toxic messages.

WHAT WE EXPECTED TO ACHIEVE

CHAT BOT	COMPARISON/ ANALYSIS OF STREAMERS	RELATION OF STREAMER AND CHAT
Twitch integrated chat bot that can detect toxic messages	Comparison of different streamers and games	Find relation between twitter messages and chat content

WHAT WE HAVE DONE

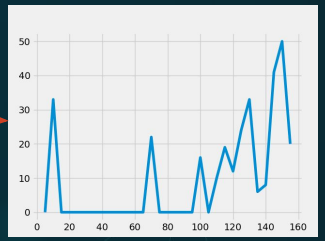
CHAT BOT	COMPARISON/ ANALYSIS OF STREAMERS	RELATION OF STREAMER AND CHAT
Twitch integrated chat bot that can detect toxic messages	Comparison of different streamers and games	Find relation between twitter messages and chat content

CHAT BOT

Unification of Data Normalization + Dictionary + FastText in a bot.

2 different approaches:

-Capture Toxicity statistics of Twitch streams with a livegraph.



-Moderate chat type depending on % toxicity.

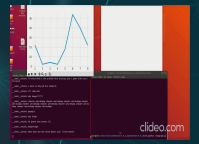
TOXICITY LIVEGRAPH

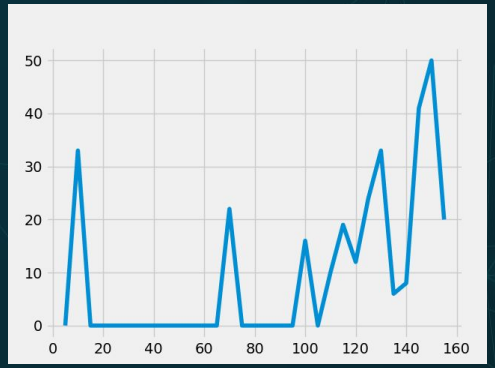
Function that calculates the toxicity of the last messages in a certain period time (each 30 seconds in our case)

Toxicity = toxic messages / total messages

The script can modulate the period time

Generate a real time graph that represents the toxicity collected.





BOT AS MODERATOR

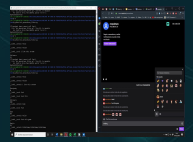
We have implemented the toxicity computation in another level for twitch streamers.

- If (in a certain time window of 30s) toxicity raises up to more than 20% → change **chat mode**

Everybody can chat → Followers only → Subscribers only → Emotes only

>20% >20% >20%

- If in the time window toxicity stays below 20% the chat mode changes to a lower level



ANALYSIS OF STREAMINGS

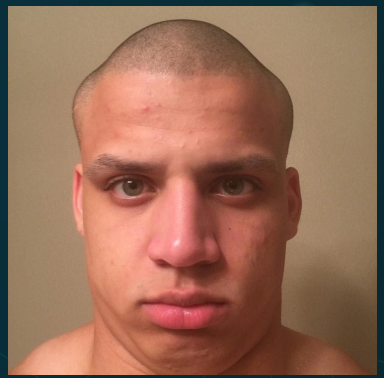
LOLTYLER1

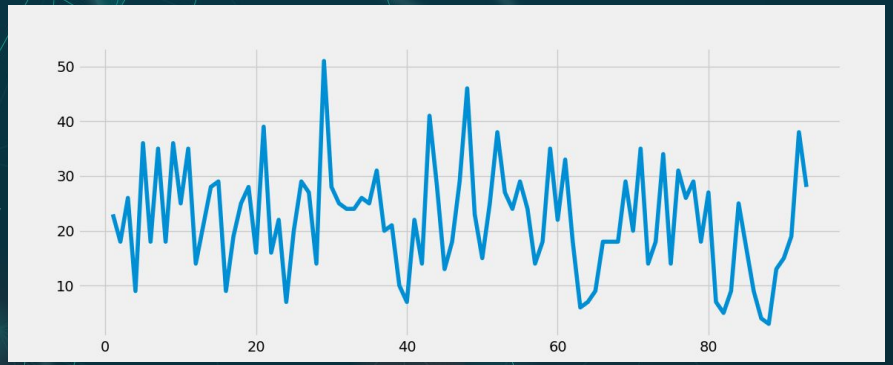
American Twitch Streamer - 2M followers

LOL - “The Most Toxic Player in North America”

19k Average Viewers

Chat mode: Only Followers





ANALYSIS OF STREAMINGS

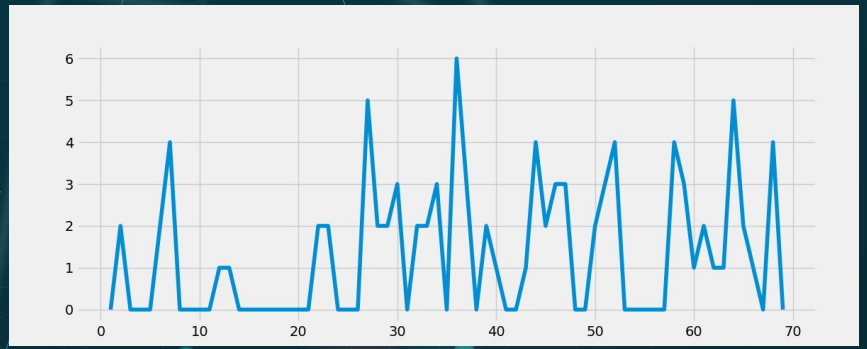
YOGCAST

CHARITY STREAM

Minecraft

25k Average Viewers

Chat mode: Only Followers



ANALYSIS OF STREAMINGS

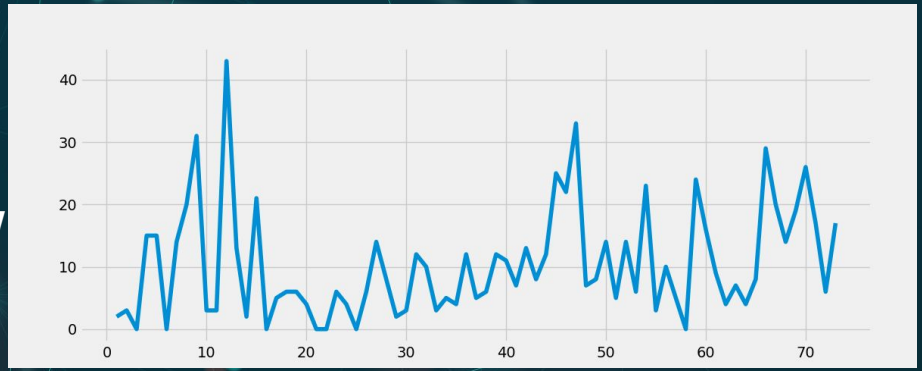
DREAMLEAGUE (DOTA 2)

Dreamleague Competition Streaming - English

DOTA 2

Chat Mode: ONLY FOLLOWERS

2k average viewers



%
Toxicity

#Intervals of 30s

ANALYSIS OF STREAMINGS

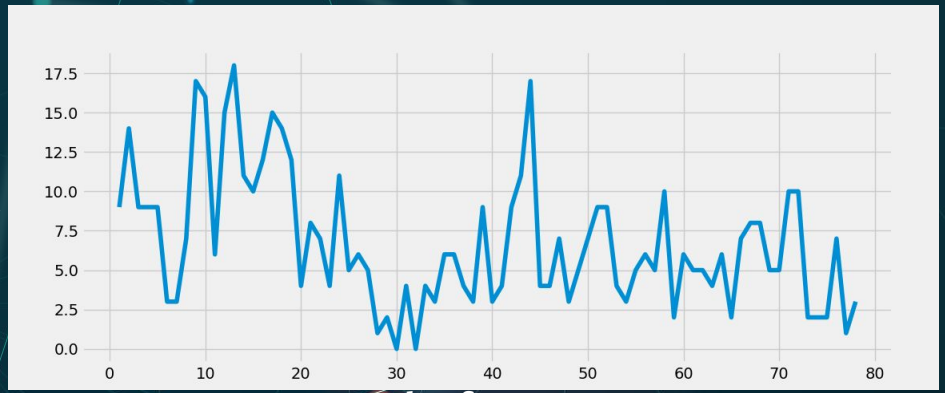
ESL (CSGO)

CSGO Pro League FInals Streaming - English

CS:GO

Chat mode - SLOW MODE (everyone can write every 5 seconds)

60k Average Viewers



% Toxicity

#Intervals of 30s

PROBLEMS DURING PROCESS

- How to label data? Criteria?
- Data is valid for our problem?
- Context of the message and how to deal with it
- How much data?
- Should we grab emotes?

FUTURE IMPROVEMENTS

- Context should be taken into account for labelling
- ML Model should also be able to understand the context
- Data labelling with more people labelling each instance independently
- More data
- More normalization? Emojis?

CONCLUSIONS

- Good results even with low volume of data
- Difficult to obtain data from Twitch due to its API
- Easy applications into Twitch thanks to its API - Chat Bots
- The type of games and community around the streamer can explain easily the toxicity of the chat

REFERENCES

- www.noswearing.com/dictionary
- https://hatebase.org/
- https://fasttext.cc/docs/en/supervised-models.html
- https://pythonprogramming.net/live-graphs-matplotlib-tutorial/
- https://dev.twitch.tv/docs/irc

Toxicity in twitch

Thank you!

Lluís Hernández
Sergi Olives
Eloi Yerpes

#SomTalaiòticsin'Eloi
Universitat Pompeu Fabra
Novembre 2019