Inappropriate behavior detection in YouTube comments

Ivanov Mikhail

May 2024

Abstract

The work shows several approaches to detect offensive/toxic behaviour in comments provided by users on YouTube platform, and distinguish those from meaningful or thoughtful. That said, the key idea is to find some pattern in such way of acting hostile to other users.
The project is placed here https://github.com/Mishquad/nlp_project

1 Introduction

This research focuses on a crucial task: distinguishing between two categories in Russian-language comments on YouTube: inappropriate (bots/trolls) and genuine/thoughtful comments.

Bots, essentially automated systems, often serve specific agendas, such as promoting advertising or targeted propaganda. Trolls, on the other hand, are individuals who aim to provoke hatred and pointless arguments within comment sections, typically with messages unrelated to the video topic.

Identifying bots is relatively straightforward—they operate based on predefined algorithms with some obvious text-pattern. However, detecting trolls presents a more complex challenge due to the nuanced nature of human behavior, making it difficult to formalize.

Comment sections across various platforms frequently encounter disruptive users, whether intentional or unwitting, who contribute to the spread of hatred or clutter the information space. These users, in turn, can be broadly categorized into two groups: individuals acting independently or at the behest of others, and automated programs known as bots, which are becoming increasingly sophisticated with advancements in text generation technology.

This research proposes multiple approaches to filter out noise (such as bots and trolls) from genuine user interactions in YouTube comments. By employing these methods, author aims to contribute to the development of more effective moderation strategies, fostering a healthier online environment where genuine discussions can exist without interference from disruptive elements.

As for object of the study, the choice fell on Yuri Dud's YouTube channel with interview of russian enterpreneur Oleg Tinkov. The video has 27 million views as of 01.05.2024 and 95 thousand comments, which is supposedly enough to reflect ongoing sentiment and text-patterns.

1.1 Team

Ivanov Mikhail prepared this document and is the author of the report.

2 Related Work

Previous studies offer valuable insights into this domain. [Skowronski (2019)] provides a foundational perspective, proposing a program utilizing supervised learning on Reddit data to detect bots and trolls. [Varol et al. (2017)] conducted a comprehensive analysis using classical machine learning techniques, achieving promising results in bot detection on Twitter. Their study incorporates features like Random Forests and visualizations through t-SNE projections, enhancing understanding and visualization of bot behaviors.

In a similar vein, [Chu et al. (2012)] introduced entropy calculations to discern human from robotic behavior, complementing traditional machine learning approaches. Meanwhile, [Kudugunta and Ferrara (2018) and Mazza et al. (2019)] explored neural network architectures to tackle similar challenges, showcasing the diverse methodologies employed in this field.

However, the absence of labeled data in our context requires innovative approaches. To address this, we propose utilizing topic modeling and Gaussian Mixture Models (GMM) with embeddings from ruBERT—a Russian language model—to identify and categorize users. This methodology draws inspiration from [Gilbert (2019) and Kong (2019)], who leveraged topic modeling, particularly Latent Dirichlet Allocation (LDA), to uncover hidden patterns in textual data.

Additionally, sentiment analysis emerges as a complementary tool for identifying disruptive behavior. Works such as [Ю.В.Рубцова (2014) and Rogers et al. (2018)] provided annotated data for sentiment evaluation, offering valuable resources for detecting bots and trolls based on emotional cues.

In light of these insights, our research aims to close the gap between theoretical understanding and practical application, proposing novel methodologies tailored to the unique challenges of Russian-language YouTube comments. By leveraging advanced techniques and drawing upon the rich tapestry of previous research, we aspire to foster a more constructive and harmonious online environment.

3 Model Description

3.1 Latent Dirichlet allocation

To analyze the corpus with comments, topic modeling is performed using latent Dirichlet allocation from the GenSim package.

3.2 Gaussian mixture models (GMM) and K-means for RuBert embeddings

The second text clustering method is based on K-Means and GMM, which is used for embeddings obtained with RuBert for each comment. Also, after preprocessing (see section 4), only the first 120 words were used in each document. This condition is due to the fact that RuBert has a limited size of the input sequence.

As a vector representation for GMM (and K-means) for a separate comment, the average embedding of RuBert from him was used.

4 Dataset

The data will be comments from YouTube channel "вДудь". On these videos, the journalist interviews famous actors, politicians, etc. The choice fell on this channel for the following reasons:

1. High popularity and as a result a lot of comments.
2. The blogger and his guests have a somewhat outrageous character, which can attract a large number of ill-wishers. It is expected that this may contribute increase in bots and trolls in the comments to this video.

For the analysis of comments, a blog dedicated to interviewing the founder of Tinkoff Bank was chosen. A program that reads data and puts it in a Tab. 1 was implemented.

Field name	Description
author_id	Commenter's channel ID
author_url	Address (URL) of the channel of the author of the comment
author_name	Author name
text	Text with comment
reply_count	Number of replies to a comment in a thread
top_level	Depth of comments
publishedAt	Publication date
updateAt	Post update date
likeCount	Number of likes per comment

Table 1: Description of the table with data from YouTube

To build a corpus with comments, the following actions are performed:

1. Calculates for each author the number of comments that he wrote underthe video.
2. Comments of rare users (commented less than 3 times) are discarded.
3. Delete stop words (English and Russian).
4. Lemmatization (With SpyCy library, file "ru_core_news_sm")

5 Experiments

5.1 LDA

A visual representation of the word distribution for each topic is shown in Fig. 1.

The class with noise comments was selected after looking at the distribution of words in themes (See Fig. 1). Let’s randomly select 10 comments from topic number 4. It looks like a topic with information noise and politics-related .

```
@АНИКС кто сказал, что война бы началась? Кто бы ее начал? Вы что-то перевираете...
------------------------------------------
@Снежков а я не сомневаюсь бот...иди проспись...
------------------------------------------
Папе наверное надоело стоять в очереди за колбасой ......
-----------------------------------------Олег, ничем не отличается
от этих коммуняк. Была какая либо власть, набросился
с...
------------------------------------------
Мы перенесли тогда в 2015 EBV Virus-- это когда иммунитет полностью на нуле, опа... ---------------------
---------------------
Олег плохо за карму борешься - лукавишь, приписываешь себе заслугу, что ещё в пе...
-----------------------------------------согласен, поэтому СВО необходимо, чтобы вылечить эту раковую
опухоль Украины, на...
------------------------------------------
Ну,не всем-же иметь таких крутых воров друзей,как у тебя...
------------------------------------------
@elm tyu как ничего не решают ? Ведь ваши люди поддерживают войну!? Даже родстве... --------------
----------------------------
@OffvaniaЯ ЖИВУ В СВОБОДНОЙ ПРЕКРАСНОЙ РОССИИ!!! ПРЕДАТЕЛЯМ НЕТ МЕСТА В НАШЕЙ ВЕ...
------------------------------------------
```

5.2 t-SNE and GMM

t-SNE is used to visualize clusters obtained with GMM. On the Fig. 4 can be seen that part of the data has peeled off from the central cloud, but GMM does not separate these points in any way. A schematic workflow is shown on Fig. 3

The choice of a class with noise comments was chosen after viewing the classified texts. Next, we print 10 randomly selected lines from text that can be considered spam.

```
дисс на og buda!. <a href="https://youtu.be/WabZ3hSz2Uo">https://youtu.be/WabZ3h...
------------------------------------------
```

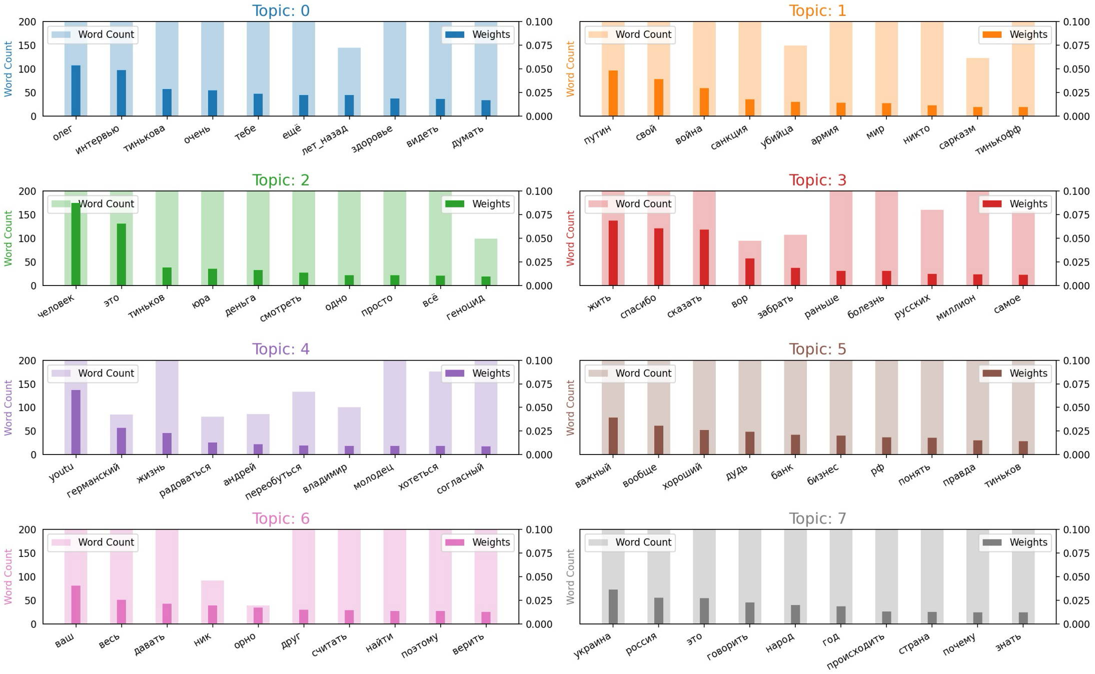

Figure 1: Distribution of words in topics

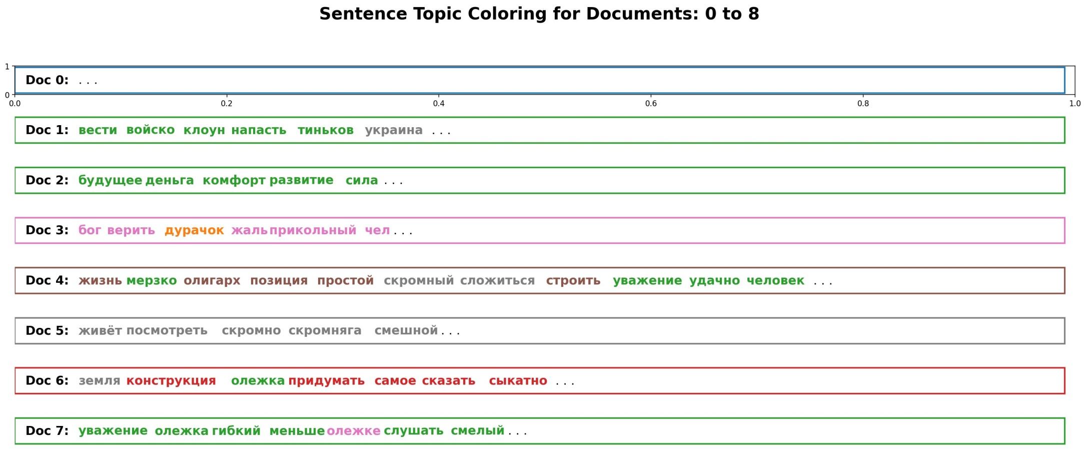

Figure 2: Some documents

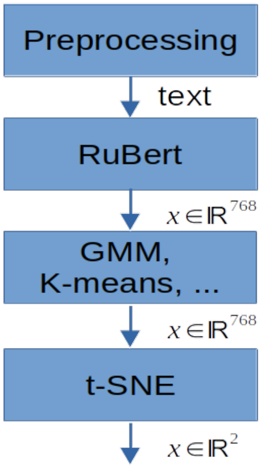

Figure 3: workflow

```
  То самое Младше !7 в Нике ...
------------------------------------------
Реаниматолог о важности кислорода и тканевого дыхания, о роли Синтезита в этих п...
------------------------------------------
От каких вопросов Украинцам у них происходит сбой в программе?На что они не смог...
-----------------------------------------https://youtu.be/JoYL1MZmXbs вот...
------------------------------------------
<b>Когда Закончится Война ..</b><br><b><a href="https://youtu.be/DfOOS2F1F44">...
-----------------------------------------С0чнble Шк0дницЫ в Нukke
<br>
<br>
<br>
<br> рассеивает солнечный свет, чт... -------------------------
----------------это наконец здесь
https://youtu.be/7RTfXNvy8ck...
------------------------------------------
<b>Тоже рады что Юра вернулся</b><br><a href="https://youtu.be/jk2t5ucbJzc">http...
-----------------------------------------peskov is a coward and a murderertoo! I don’t think he is a wonderful
person aft... ------------------------------------------
```

5.3 t-SNE и K-means

For K-means similar Fig.5 is built previous t-SNE projection, but with markup after K-means.

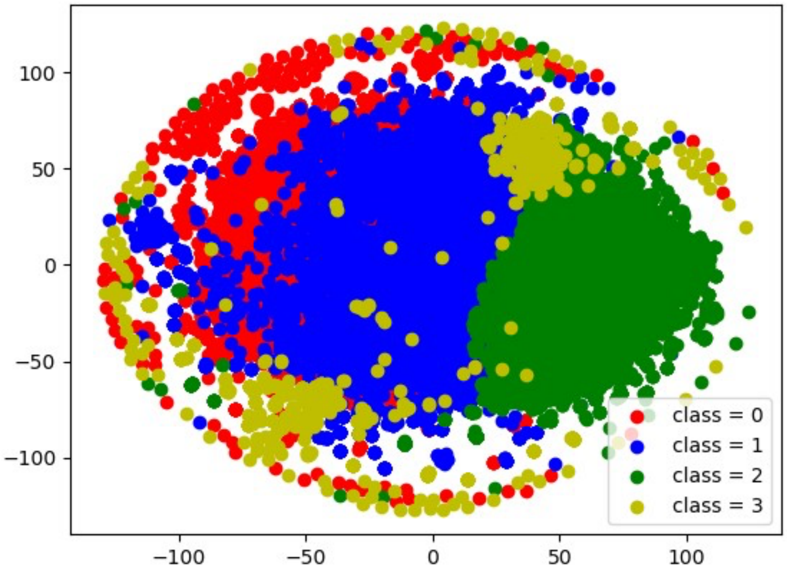

Figure 4: t-SNE и GMM (Noise comments: class 3)

The choice of a class with noise comments was chosen after viewing the classified texts. Eight randomly selected points from informational noise:

Z...

Если бы мы не вели войска в Украину она бы напала на нас а Тиньков он клоун...

Если бы мы не вели войска в Украину она бы напала на нас а Тиньков он клоун...

В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее....

Если бы мы не вели войска в Украину она бы напала на нас а Тиньков он клоун...

В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее.... ---

В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее.... --

В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее....

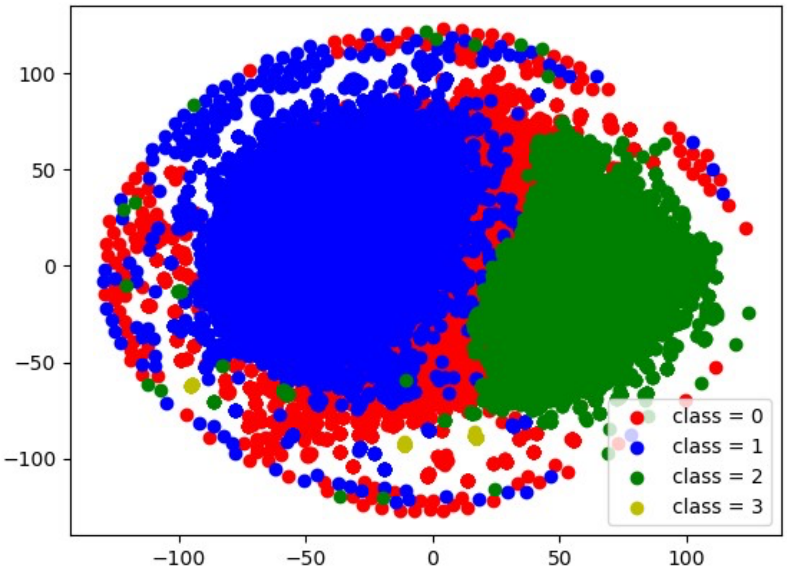

Figure 5: t-SNE и K-means (Noise comments: class 3)

5.4 t-SNE with points on the periphery

It was previously noted that after projecting data onto a two-dimensional plane using t-SNE, some of the points are peeled off (Fig.4). Perhaps these points represent some irregular data, in other words outliers. These emissions may be related to the informational noise of trolls and bots. Separate points on the periphery. To do this, we will take points only outside the circle of some radius (see Fig 6).

We also present 10 randomly selected lines from the text classified in this way:

Не важно что и как в Украине, не важно вообще нечего, что там происходило и прои... ---P...

<a href="https://youtu.be/Krcr9IQ0Q_w?t=9039">https://youtu.be/Krcr9IQ0Q_w?t=903... --

Eres un ı´dolo <a href="http://ayumi.monster/">Ayumi.Monster</a><br><br>mejores.<...

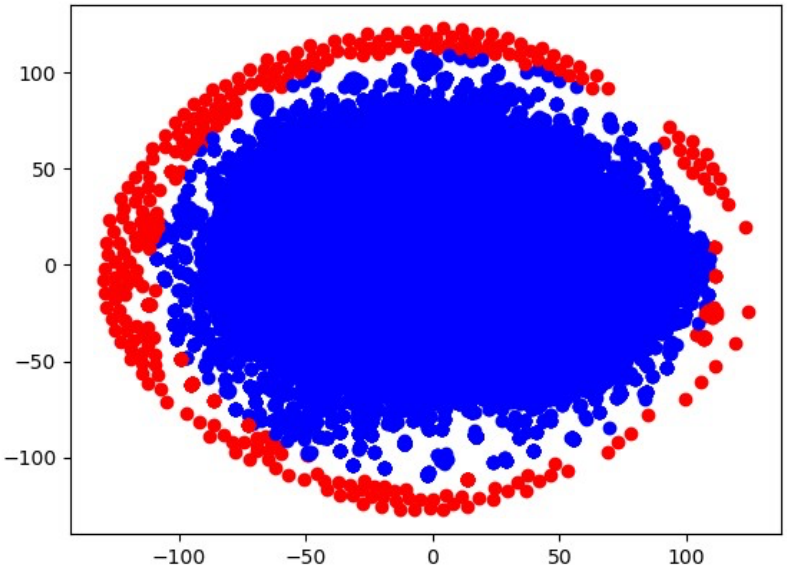

Figure 6: t-SNE and selection of points on the periphery of a circle

```
-------------------------------------------
<a href="https://www.youtube.com/watch?v=Y3hNyPKuEUU&amp;ab_channel=NEMAGIA">htt... ------------
-------------------------------
<a href="https://www.youtube.com/watch?v=Y3hNyPKuEUU&amp;ab_channel=NEMAGIA">htt...
------------------------------------------То самое Младше !8 в
  нuke ...
------------------------------------------То самое Младше !8 в
  нuke ...
------------------------------------------Каратели.
<br>Неужели мозги вы заржавили, <br>Дорогой
украинский сосед?
<br>Вы н...
-------------------------------------------
Топ...
```

5.5 Experiment Setup

The Fig. 3 shows a general scheme for clustering and visualizing data using RuBert, GMM (K-means), t-SNE.

The following parameters have been set for LDA:

- Number of topics: 8
- Number of iterations: 100
- α: symmetrical
- Number of passes through the body during training:10
- Number of documents to be used in each training block: 10

6 Results

The Tab. 2 shows the number of comments that were filtered by one of the four methods. Only the comments of authors who wrote more than two times participated in the clustering. Similar statistics for authors only are presented in the Tab. 3.

Fig. 1-5 shows the clustering results.

models	class_size
total	22661
GMM	976
out_circle	436
kmeans	413
lda	2185

Table 2: General statistics on noisy comments

models	number_of_authors
total	4267
GMM	319
out_circle	58
kmeans	255
lda	1455

Table 3: General statistics for noisy authors

7 Conclusion

Three approaches were used to identify troll and bot comments. The first is topic modeling with LDA. The second method of parsing messages is based on the use of embeddings obtained from RuBert. Further, for this vector representation, the GMM and K-means clustering methods were used. Based on the hand review of the data for the corresponding clusters, a conclusion was made about the class number to which the noise messages correspond. The third approach is tdistributed stochastic neighbor nesting (t-SNE) applied to RuBert embeddings. Noise comments were considered points lying on the periphery of a circle of a given radius.

As a result, the analysis of texts that were marked as noise (with bots and trolls), we can conclude that thematic modeling showed the worst. There is a lot of information in his texts that does not look like the work of trolls and bots.

Methods based on RuBert embeddings showed very good results. Approach based on the identification of noise points at the t-SNE boundary performance showed good results. This method does not require additional analysis of classes like all other methods. The only thing he selected is less than all the points (objects) for the noise class. This is due to the choice of a large circle radius. The following experiments were not successful in the work:

- Detect GMM points that lie far from the centers of Gaussians of each class.
- Build a classification model on labeled data.

References

[Chu et al., 2012] Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*.

[Ю.В.Рубцова, 2014] Ю.В.Рубцова (2014). Построение корпуса текстов для настройки тонового классификатора. *Программные продукты и системы*.

[Gilani et al., 2017] Gilani, Z., Farahbakhsh, R., Tyson, G., and Crowcroft, L. W. J. (2017). Of bots and humans (on twitter). *Conference: 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[Varol et al., 2017] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv*.

[Rogers et al., 2018] Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., and Gribov, A. (2018). Rusentiment: An enriched sentiment analysis dataset for social media in russian. *Proceedings of COLING*.

[Kudugunta and Ferrara, 2018] Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*.

[Kong, 2019] Kong, B. (2019). Analysing russian trolls via nlp tools. *The Australian National University*.

[Gilbert, 2019] Gilbert, E. C. E. (2019). Hybrid approaches to detect comments violating macro norms on reddit. *https://arxiv.org/abs/1904.03596*.

[Mazza et al., 2019] Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., and Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. *WebSci*.

[Pierre, 2019] Pierre, S. (2019). Russian troll tweets: Classification using bert. *towardsdatascience.com*.

[Rogers et al., 2018] Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., and Gribov, A. (2018). Rusentiment: An enriched sentiment analysis dataset for social media in russian. *Proceedings of COLING*.

[Skowronski, 2019] Skowronski, J. (2019). Identifying trolls and bots on reddit with machine learning. *towardsdatascience.com*.

[Alsmadi and O'Brien, 2020] Alsmadi, I. and O'Brien, M. J. (2020). How many bots in russian troll tweets? *Information Processing and Management*.

[Tardelli et al., 2022] Tardelli, S., Avvenuti, M., Tesconi, M., and Cresci, S. (2022). Detecting inorganic financial campaigns on twitter. *Information Systems*.