Detect trolls and bots in YouTube comments

Lebedev Maxim

December 2022

Abstract

The work presents several approaches for separating informational noise (bots, trolls, etc.) in the comments under YouTube videos from other information. The project is located here: `https://github.com/mvlebedev/ods_nlp_huawei_fall_2022`.

1 Introduction

This work is devoted to the development of an approach for definitions of bots and trolls in Russian-language comments on the YouTube service.

In most cases, bots mean automated systems that promote a certain agenda in the interests of customers. It can be advertising or targeted propaganda. On the other hand, trolls are mostly people who try to use messages in the comments (in most cases not related to the topic of the material being commented) to provoke hatred in the audience and force them to arrange a useless skirmish. Detecting trolls is a more difficult task, because it is more difficult to formalize.

All services in which it is possible to write comments collide with the problem of unscrupulous users who intentionally or unintentionally incite hatred (advertise and litter the information space).

Such users can be conditionally divided into two groups. The first group is living people who do this on their own or someone else's initiative. Second group are automated programs (bots) that are based on comment generators. With the development of text generation technology (such as for example GPT-2), such programs are getting better and better every day. People can also be divided into two groups: those who do it on their own and those who do it at the request of employers.

The work presents several approaches for separating informational noise (bots, trolls, etc.) in the comments under YouTube videos from other information.

1.1 Team

Lebedev Maxim prepared this document and is the author of the report.

2 Related Work

Among the works devoted to the topic under consideration, one can single out a couple of articles on the Internet [Skowronski, 2019]. This work considers the presented problem in general, and also gives a specific way to implement a program to detect bots and trolls. It uses supervised learning on already labeled data taken from Reddit. This work uses various metadata about users, as well as numerical features for the text, characterizing its diversity. The model is a decision tree.

A number of papers [Varol et al., 2017, Alsmadi and O'Brien, 2020], [Tardelli et al., 2022] also use classical machine learning approaches.

In [Varol et al., 2017] presented a large-scale study on manually labeled data from Twtter. To solve the problem of classifying bots, classical machine learning methods were used: Random Forests, AdaBoost, Logistic Regression and Decision Tree classifiers. The best result was obtained using Random Forests. Also in this work, a t-SNE (T-Distributed Stochastic Neighbor Embedding) projection was built for the most important features to visualize classes on a two-dimensional plane. In addition, the BotOrNot program was created to detect bots on twitter. [Gilani et al., 2017] compared the quality of detection of bots by this program with these marked people. In the article [Chu et al., 2012], along with the classical approach (Random Forest), entropy is calculated for time counts of service calls. High entropy, according to the authors, characterizes a human while regular behavior (as a result is low entropy) are characteristic of robots.

Other works use neural networks [Kudugunta and Ferrara, 2018], [Mazza et al., 2019, Pierre, 2019].

In the conditions of the problem posed in this work, there are no labeled data. Therefore, it is proposed at the first stage to understand what these users are. For this, thematic modeling and Gaussian Mixture Model (GMM) for embeddings from RuBert will be used. The papers [Gilbert, 2019, Kong, 2019] also formulated an approach using topic modeling, for which Latent Dirichlet Allocation (LDA) is used.

In the work [Ю.В.Рубцова, 2014] the marked data for the evaluation of the sentiment of the texts was obtained. [Rogers et al., 2018] contains tagged message data from the VKontakte social network. On the basis of this corpus, the Dostoevsky library was developed to evaluate the tone of the text. Sentiment analysis can be used as an additional tool to find bots and trolls.

3 Model Description

3.1 Latent Dirichlet allocation

To analyze the corpus with comments, thematic 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 (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 blogger 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.

Was implemented, a program that reads data and puts it in a Tab. 1.

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 under the 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.

```
 То самое Младше !7 в Нuke ...
-------------------------------------------
@NekoSoul EngMafaka Революція Гідності це дуже крута штука, і якби її не було, т...
-------------------------------------------
ще не вмерла?...
-------------------------------------------
Дудяха хайпанул...
-------------------------------------------
Вышибайте их постоновщика.. В носке!!!...
-------------------------------------------
В топ чтобы все увидели!...
-------------------------------------------
Огонище...
-------------------------------------------
@артем гомозов а вы откуда знаете где базировалась Украина? Есть пруфы? Я могу с...
-------------------------------------------
Шныри...
-------------------------------------------
 То самое Младше !8 в нuke ...
-------------------------------------------
```

5.2 t-SNE и GMM

t-SNE is used to visualize clusters obtained with GMM. On the Fig. 4 it 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.

```
ск0льк0 6удет продолжаться конфликт России и Украины?И что нужно для его пекраще...
-------------------------------------------
https://www.youtube.com/watch?v=Y3hNyPKuEUU&ab_channel=NEMAGIA...
```

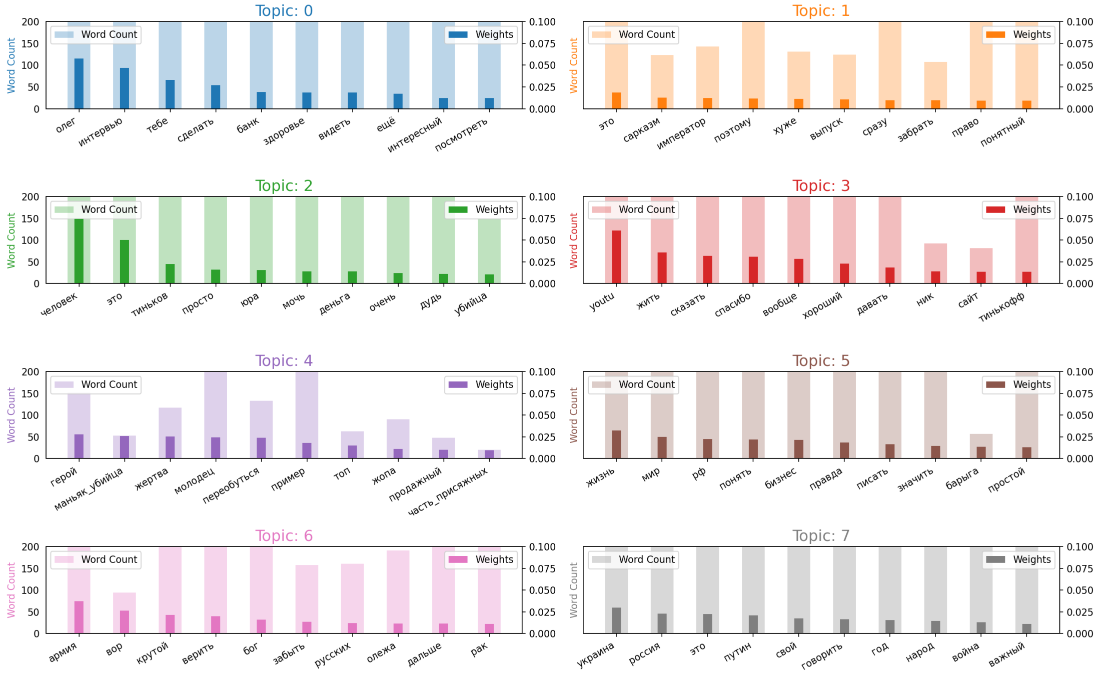

Figure 1: Distribution of words in topics

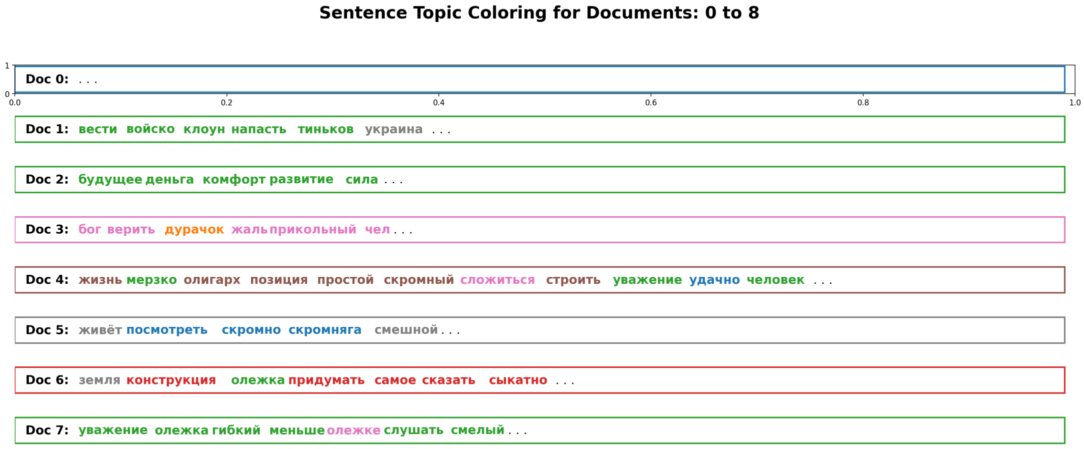

Figure 2: Some documents

```
---------------------------------------------
 То самое Младше !8 в нuke ...
---------------------------------------------
```

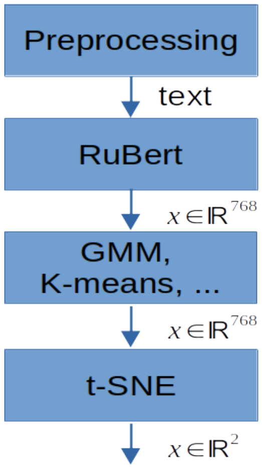

Figure 3: workflow

```
<a href="https://www.youtube.com/watch?v=Y3hNyPKuEUU&amp;ab_channel=NEMAGIA">htt...
-------------------------------------------
@Anastasiia Savarovskaya я все помню...
-------------------------------------------
 То самое Младше !8 в нuke ...
-------------------------------------------
@aligagator bezze а ты там был когда горело???...
-------------------------------------------
 То самое Младше !8 в нuke ...
-------------------------------------------
@JnandGUK по однобоким фактам....
-------------------------------------------
I do not believe that Russia can win this war. Even the president of Belarus, a ...
-------------------------------------------
```

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.

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

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

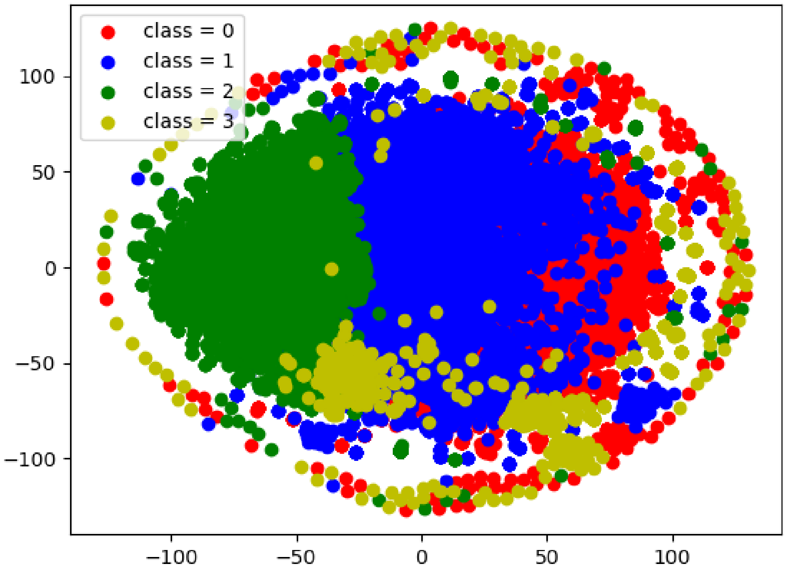

Figure 4: t-SNE и GMM (Noise comments: class 3)

```
В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее....
-------------------------------------------
Z...
-------------------------------------------
Z...
-------------------------------------------
Если бы мы не вели войска в Украину она бы напала на нас а Тиньков он клоун...
-------------------------------------------
Z...
-------------------------------------------
Z...
-------------------------------------------
В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее....
-------------------------------------------
В деньгах силах. Есть деньги, есть комфорт, есть развитие, есть будущее....
-------------------------------------------
```

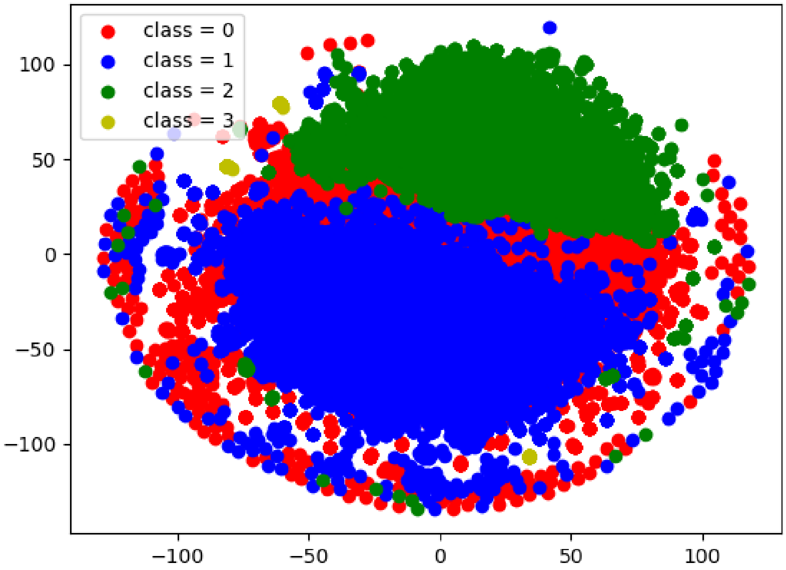

Figure 5: t-SNE и K-means (Noise comments: class 3)

5.4 t-SNE with dots on the periphery

It was previously noted that after projecting data onto a two-dimensional plane using t-SNE, some of the points 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:

```
Ghana Socialist Leader: We Know Who Are Real Enemies Are and It’s Not Russia
yo...
-------------------------------------------
<a href="https://www.youtube.com/watch?v=Y3hNyPKuEUU&amp;ab_channel=NEMAGIA">htt...
-------------------------------------------
Не важно что и как в Украине, не важно вообще нечего, что там происходило и прои...
-------------------------------------------
Топ...
-------------------------------------------
 То самое Младше !8 в нuke ...
```

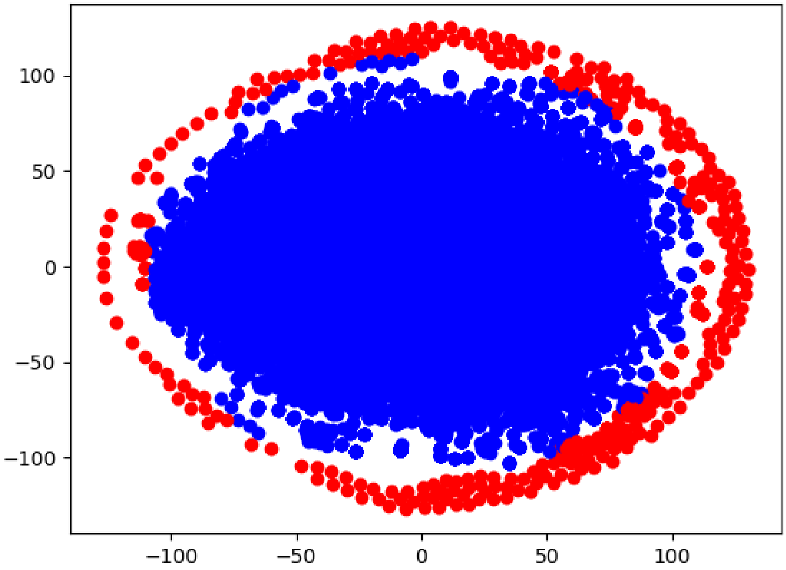

Figure 6: t-SNE and selection of points on the periphery of a circle

```
-------------------------------------------
Все русские патриоты поддерживают специальную военную операцию.

Против только ...
-------------------------------------------
 То самое Младше !8 в huke ...
-------------------------------------------
<b>Не Дудь&#39;</b><br><a href="https://youtu.be/ItFVgpC_YTM">https://youtu.be/I...
-------------------------------------------
Путь ВЕЛИКИХ БЕЖЕНЦЕВ!!!

Путь ВЕЛИКИХ БЕЖЕНЦЕВ!!!

Путь ВЕЛИКИХ БЕЖЕНЦЕВ!!!
...
-------------------------------------------
 То самое Младше !8 в huke ...
-------------------------------------------
```

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 table 3.

Fig. 1-5 shows the clustering results.

models	class_size
total	22661
GMM	976
out_circle	503
kmeans	413
lda	2056

Table 2: General statistics on noisy comments

models	number_of_authors
total	4267
GMM	319
out_circle	61
kmeans	255
lda	1281

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 t-distributed 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

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

[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.*

[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).*

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

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

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

[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.*

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

[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.*

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