A BERT Approach to YouTube Comment Spam Detection as a Pairwise Sequence Classification Task

Samuel Morris **Tony Lay**

1 Background and Motivation

YouTube is a highly-renowned video-sharing platform with more than 2 billion active monthly users worldwide (YouTube, 2023). Users are able to create a 'YouTube channel' which allows them to upload videos and accumulate video views as well as recurrent viewers known as 'subscribers'. Due to the increasingly large number of users in recent years, there is an inherent monetary incentive to grow and maintain a YouTube channel by creating videos, gaining subscribers and amassing a large audience - resulting in the potential to establish a strong online presence, generate ad revenue, sell premium member-only channel features and more (YouTube, 2021). This has resulted in a massive growth in the development of automated bots capable of large-scale deployment of unsolicited advertisements (spam) across numerous large channels, in order to promote their own channels or malicious external links. It is clear that the presence of such comments are obstructive, and it damages the experience of normal users as well as the reputation of the channel being posted on. The mass distribution of spam is also a violation of clause 5 of 'Permissions and Restrictions' in YouTube's Terms of Service (YouTube, 2022). YouTube have deployed a handful of procedures to tackle this issue such as implementing detection of hyperlinks, detecting repetitive posting of comments and using their own machine learning models (Peters, 2022). However it's obvious that these methods can be improved on as bots have managed to avoid YouTube's restrictions by creating new accounts as well as techniques which exploits human curiosity, e.g. "Don't click on my channel" as example of reverse psychology and "I'm better than this channel", an example of 'rage-baiting'. We propose the use of NLU techniques to tackle these issues and form the basis of spam detection in YouTube comments.

2 Problem Statement

We refer to comments as spam if the intent of the posted comment is to promote links to websites or videos that are contextually irrelevant to the original video that they posted on. The problem we wish to address is the detection of spam in the comments section for every video. The information that we will use for our input are comments fetched from videos found on the 'Trending' page as there is a high abundance of bot comments here. Comments are sentences composed of words and punctuation - we intend to keep the punctuation as we believe that it may be beneficial to differentiating between ham and spam. We will also consider the use of channel names, as examples like 'Don't click on me' as a channel name is a strong indication of spam. A completely separate input is also passed into our model which we will call the 'context' - this is a generalisation of the context of the video in the form of a sequence of key words. The context can be derived from the captions, title, tags, and description of the video. The point of having a context input is to provide a heuristic towards spam/ham classification by comparing the relevance of the comment to the sequence of context words, which aims to represent the context of the video. If the comment is highly relevant to the video, it is more likely to be ham and vice versa. Therefore, the underlying NLU task of our spam detection problem is a pairwise sequence classification, and the output of our model will be 'spam' or 'ham'.

3 Related Work

There exist multiple relevant solutions similar to our chosen problem, the identification of spam is relevant to many areas, one solution by Rădulescu et al. (2014) to detect spam comments on content sharing applications using natural language techniques such as part of speech (POS) taggers, and evaluation with three different classifiers; Naive

Bayes, Support Vector Machines (SVM), and Decision Trees (DT), shows it is possible to detect spam comments based on certain features that usually appear in them. Another similar solution found N-Grams to be effective in the detection of spam comments on YouTube (Aiyar and Shetty, 2018), this approach also made use of conventional machine learning approaches like the previous solution.

At its base our problem is a pairwise sequence classification problem, current research has found that Transformers are the state of the art solution to these Natural Language problems (Wolf et al., 2020). The BERT (Devlin et al., 2018) masked language model is capable of learning effective latent representations, shown to provide increased benefits to multiple natural language tasks (González-Carvajal and Garrido-Merchán, 2020). To support the pairwise sequence classification task, a topic modelling approach will be employed in order to retrieve a sequence of context words to compare to. State of the art solutions to topic modelling include the very popular Latent Dirichlet Allocation (LDA) (Blei et al., 2003) modelling library, as well as a more recent technique known as Additive Regularization on Topic Models (BigARTM) - a non-Bayesian regularized model (Vorontsov et al., 2015). We aim to investigate the use of both approaches.

The aim for our project is to develop a solution to the known problem of spam detection and compare it to a previously reported solution. We are going to compare it to the solution by Aiyar and Shetty (2018), as this task is also focused on YouTube comments and it is a high performing well reported solution. Their solution does not make use of pairwise sequence classification, we hope that including this heuristic will provide benefits in the identification of common spam and spam irrelevant to the video content, potentially increasing recall of spam comments.

4 Datasets and Evaluation Resources

Searching online databases we could not find an existing database that satisfied the need of having labels for distinguishing between 'spam' and 'ham' as well as video context, so we will need to gather the data ourselves. The plan to create our dataset is to make use of the YouTube Data API to gather comments from the YouTube website, using this we can extract top-level comments and their replies. We will also filter these so we only have English comments, as not only does this reduce the complexity of dealing with multiple languages, but we need to manually label the comments. The comments of our dataset will be labeled 'spam' or 'ham'. To obtain the context, which summarises the video content the comments are posted on, the YouTube Data API provides ways to obtain items which we could use for this context, i.e. video titles, tags, captions, descriptions. We aim to create a dataset containing 5000 comments with their relevant video contexts, we believe this is reasonable given the time frame of the project and will provide enough data to train a model.

Data gathering will be completed using Python to obtain the raw data via the API then automatically filter out the unnecessary part of the API response and then the non-English comments. We then plan to create a simple program to allow us to label the comments into their respective classes. This process will provide us with a substantial dataset which will be fit to our needs, allowing us to evaluate our proposed solution to a high degree of confidence for the task presented.

5 Proposed Activities

Activity	Any comments	Duration	Lead
Dataset Generation	Gathering and labeling of YouTube comments via the YouTube Data API.	1 Week	SM
Topic Modelling	Deriving context from video captions, titles, descriptions, and tags	1 Week	TL
Creating Neural Network	Using BERT as a baseline for the neural network	2 Weeks	TL
Training & Testing	Hyper-parameter selection and analysis of the model	1 Week	SM
Code Refactoring	Fix any minor bugs, clean-up the code, improve the system if there is time left.	1 Week	SM

References

Shreyas Aiyar and Nisha P Shetty. 2018. N-Gram Assisted Youtube Spam Comment Detection. *Procedia Computer Science*, 132:174–182. International Conference on Computational Intelligence and Data Science.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Santiago González-Carvajal and Eduardo C. Garrido-Merchán. 2020. Comparing BERT against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Jay Peters. 2022. Youtube's new weapons for fighting comment spam include 24-hour bans.

Cristina Rădulescu, Mihaela Dinsoreanu, and Rodica Potolea. 2014. Identification of spam comments using natural language processing techniques. In *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 29–35.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. 2015. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 29–37.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

YouTube. 2021. How to Make Money on YouTube - YouTube Creators.

YouTube. 2022. Terms of Service.

YouTube. 2023. YouTube for Press.