UNIVERSITÀ DI TRENTO

Computational Social Sciences:

research project proposal

Understanding intense online sentiment: The Youtube Apology and related intense online discourse using NLP

A.A. 2022/2023 *Davide Vandelli*

Introduction

A "YouTube Apology" is a relatively recent term that indicates a form of crisis communication regarding an Internet *persona* (a YouTuber) that is directed at their fans, or a more general public invested in the *persona*'s online and / or real life mistakes and scandals. Often, research on the topic studies the strategies that the Internet *persona* implements to seeks forgiveness, redirecting negativity or trying to settle their online reputation; when the figure apologizing is an influencer, their careers often depends on it (*Sandlin & Gracyalny,* 2018), or if they represent a commercial service[1], it might be an opportunity to strengthen the relationship with their customers - or even obtain new ones (*Park & Choi*, 2023).

Theoretical framework

Due to the relatively new context that surrounds the dynamics of the YouTube apology, the research on the topic is still divergent, complex and holds different scopes based on the field it comes from. Many also do not implement computational methods, but a more qualitative and exploratory approach.

Regarding the scope of this research, which is to deepen the understanding of this specific form of social media interaction using computational methods, there are three main gaps to be addressed.

1. What aspects of a YouTube video comment section constitute a *social* interaction, and which do not?

[1] Even though influencers can become the spokesperson for a company or a commercial service, their brand is usually disconnected from organizations such as Nivea's apology (Tsang, 2017), who makes their communication much less personal and more *ad hoc* for the scandal involved.

2. Which factors predict viewers' reactions? And which belong to the viewer, the online platform or the *apologizer*?

3. What makes online discourse *intense*?

The first two questions are by far the most complex, but their answers would also represent most of the phenomenon.

Respectively to the gap questions, I gathered some of the interpretations available.

In 2016, Hall investigated the difference between social media and social interaction: only 2% of social interaction took place through social media in a 5 day study with 54 young participants. Since then, YouTube has changed much of its features, but due to its hybrid nature (it's both online video sharing and social media) it may be a stretch to consider it a social setting, where typical social dynamics of interactions occur. However, research by Bou-Franch *et al.* (2012) had examined the conversations that occur on the platform, effectively interpreting them as social interaction. Other researchers have effectively used the term "parasocial" in front of interaction to highlight the dubious nature of social media interactions, such as Kurtin *et al.* (2018). On an operational level, we may consider the comment section to be interpreted and studied as a social interaction if the actions that take place are comparable to a conversation and reaction towards a social issue.

There is no extensive research on what factors contribute to a specific reaction from the viewer purely sociological or economic research; there is from Communication Studies (already mentioned previously) that investigate the best strategies to avoid certain reactions, or from the cognitive and emotional engagement such as Kadir *et al.*'s work (2016).

Would the age, the ethnicity, personality traits and other personal branding characteristics (other than communication strategies) contribute to explaining a public's reaction when a person is apologizing? And would they come into play from the perspective of the person viewing an apology? From a more materialistic approach, some theorize that social media is actually shaping public domains (health, education,...) of society - not just the other way around - as a constructed reality by the institutions and the people in power (*Couldry & Van Djick*, 2015). In order to attribute which factors belong to what agent is involved between the platform, the apologizer and the viewer, more research needs to be conducted, therefore leaving the second question unsettled.

Finally, the last question is approached with research from *Katz et al.*, (2004). Due to the sarcastic nature of many of the comments, utilizing research on the negativity of sarcasm online contributes to define more clearly the connotation of certain comments - more in detail whether they are positive, negative or neutral. In this research proposal I will also present an evaluation of the model's performance using a small dataset consisting of 1000 comments that I labeled as positive, negative or neutral. Based on previous research (*ibidem*), the use of non-literal language, altogether with humorous remarks that parodize the YouTuber, are labeled as negative due to the context of an apology. More information is included in the table below for specifying which attributes denote which label.

Table: features for a positive, neutral and negative comment.

	Attributes	Example
Positive	Words of affection or sympathy, positive emojis, dismissive remarks about "*haters*"	"*We miss you colleen come back when you feel okay take your time please we ll be waiting big hugs to you*"
Neutral	Comments that do not attack a person or a category, usually making a point without tone indicators.	"*Chat is this real*"
Negative	Sarcasm, parody, and explicit insults.	"*Be for real now ??? this is mental illness*"

For more obvious reasons, explicit insults towards the *apologizer* are also annotated as negative. On the other hand, lengthy comments that argue without using emotional tone indicators (upper case, emojis, punctuation, accusing sentences) are classified as neutral.

Methodology and models

Data

For the purpose of this proposal, YouTube's publicly available data is more than fit, but it could be argued that the "Apology video" format is part of phenomena observable in more than only YouTube, as recently showcased by C. Ferragni's Instagram (2023). I signed up for a *developer account* YouTube Data API v3 (YouTube Developers, 2023), in order to retrieve 12,000 comments from one specific media in July 2023. The data belongs to a video uploaded by Colleen Ballinger, also known under the stage name *Miranda Sings*. As per usual on YouTube with several apology videos[2], many viewers feel emotionally invested and compelled to express their view as comments.

Natural Language Processing with XLNET and VADER sentiment

Sentiment analysis may be tied to **rule-based** models or **neural network-based** models. Depending on the purpose and the nature of the data, different approaches can yield very different results.

Due to new slang utilized in the comments, using a lexicon-based model for the analysis may miss a significant amount of meaning, however, using neural network models may fail to capture the most recent innovations in the structure of the sentence. There are many models that are commonly deemed as fit for the purpose of being as flexible as possible. In this research proposal, the following are considered for online sentiment analysis: **XLNet** and **VADER**.

XLNet was released by Google in 2019 on a corpus of 33 billion words as a better alternative to the model BERT (*Yang et al.*, 2019).

[2] Several other examples of very similar dynamics may include: James Charles', FineBrothers', Jeffree Star's, all referenced at the end of this proposal.

VADER sentiment was developed by a smaller team of Georgia Institute of Technology Atlanta, but was specifically trained on several corpuses that had human-reviewed Internet slang (*Hutto & Gilbert*, 2014).

XLNET is one of the most common models for NLP due to its massive improvement to the field. It is often used in its pretrained version, and it blends auto-regressive (AR) and auto-encoding (AE) techniques using a permutation language modeling goal. Its neural architecture incorporates Transformer-XL and a two-stream attention mechanism. XLNet shows significant advancements compared to prior pretraining approaches across different usages.

VADER instead is a much more parsimonious approach, as it is lexicon and rule-based, developed to be a better alternative to models such as the famous SentiWordNet - it is able to run on CPU. The latter is also able to encode Western-style emoticons, which are greatly used under YouTube videos.

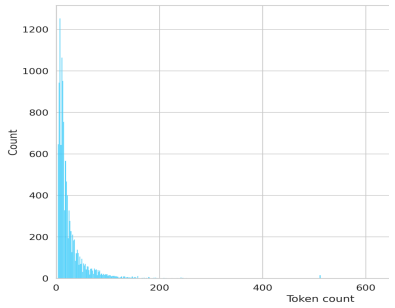
Preparing the model for analysis

With respect to the focus of this research, both models were used to evaluate how intense the reactions are towards a YouTube apology video, and towards which semantic charge they are leaning to, inside of the one-dimensional continuum from *negative* to *positive*.

Each model requires to be set and prepared for analysis, but in different manners. It is common practice to clean the data from punctuation and replace upper case letters with lower ones, but one of the two models actually uses uppercase to better understand the tone of a document.

XLNet requires the most common data cleaning techniques; furthermore, it requires GPU acceleration, as it utilizes transformers that involve complex computations. Once the pre-trained model is installed on the machine, tokenizing the data is the first step to evaluate if the prior process was done correctly. As a post-hoc analysis, it is usually required to know what is the distribution of the tokens, in order to understand what the model is working with - in the case of YouTube comments, most pieces of the data require no more than 200 tokens. In figure 1 the (exponential) distribution of the token is shown: many comments are short and intense, and the lengthier the comment is the more tokens are involved.

Fig 1. Histogram of tokens. Median = 15, standard deviation = 35.06. N = 11958.



Only a few comments go past 200 tokens, and after further inspection they are simply repeated versions of a shorter comment. As a brief qualitative description, we may say that the comments that range around 200 are long sequences that express a point of view on the matter, through points and arguments. Meanwhile, extremely short comments usually express more intense but more heuristical discourse, with emoticons and emojis as tone indicators, that either weaponize sarcasm or

declare affection and love towards the *internet persona* depicted in the video.

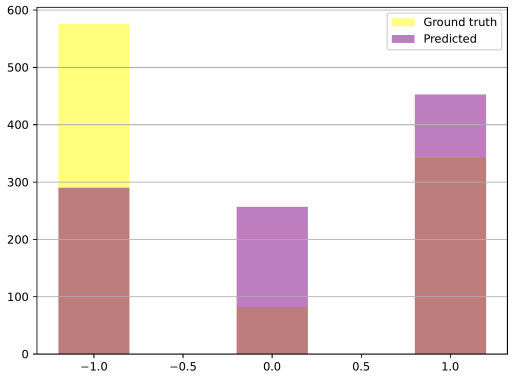
XLNet requires training for downstream tasks; when trained using the recommended hyperparameters, it may reach accuracy above 92% (*ibidem*). It is a rather complex model and its results are difficult to interpret, but it also brings much more flexibility on different domains (or social context) than other models, as it can be fine tuned with specific data, appropriately human-reviewed.

On the other hand, VADER sentiment is much more easily applicable and time-saving. Some features of the model are reported later. The latter model was originally trained by its creators on *tweets*, *Amazon reviews, Rotten Tomatoes*, and other forms of online reviews or micro-blogging corpuses. This rule-based model utilizes a *sentiment lexicon* produced from a wisdom-of-the-crowd approach, meaning that the lexicon is a gold standard with elements (e.g. words) that are deemed to be more or less intense by humans. Using VADER with its lexicon, we are able to produce one positive, one negative and one neutral score for each document, which are then turned into a *compound* score. A compound score is a normalized sentiment score ranging from {-1, 1}. The positive, negative and neutral scores are equal to the proportion of text that falls in the respective category - emojis and emoticons included. It is also sensitive to lowercase and uppercase words, as the latter usually indicate more emphasis.

Since VADER cannot be as easily trained as XLNet, I annotated a small dataset of 1,000 comments as negative (-1), neutral (0) or positive (1), after defining a few criteria to evaluate the performance of the simplest model.

Proto-findings

Figure 2. VADER sentiment data and human reviewed data. Ground truth vs Predicted values. N = 1000.



Note: in every bar, the bottom color (*brown*) represents the cases in which both VADER and my evaluation of the text are aligned.

Only VADER was applied as a first approach for this proposal, due to its highly interpretable results and prediction process. VADER predicted the meaning of the comments with 54.2% accuracy - which is 21.2 percentage points above a random classification (where each class has 33% of being selected). As I was reviewing the comments, I had noticed multitudes of sarcasm and Internet slang that probably spread on the Internet later than VADER's training. This time discrepancy may explain the many mistakes that the model in question had made regarding negative comments: half of them went unnoticed and scored as neutral, when they were likely to be sarcastic. As an example of this discrepancy, in the table below there are 5 randomly selected examples from the validation sample.

Table: VADER vs qualitative annotation.

Qualitative evaluation	*VADER Predicted*	*Text*
-1	0	GIRL REALLY?????
1	0	Idk what she did. But I forgive her
-1	0	there's literally no way this is real
-1	0	I vote this the BEST APOLOGY VIDEO of all time...
-1	-1	new worst YouTube apology vid dropped. Laura …

Every case of sarcasm and exaggeration went unnoticed. Sarcasm is usually weaponized as negative in this context. Additionally, it is not clear whether there is a solution (even for transformer-based models) for comments as short as "*GIRL REALLY?????*" to be correctly classified as indicating a negative meaning - assuming the commenter meant disappointment and shock.

Instead, when we focus on positive comments, it seems as VADER is able to classify them better than other categories. This could be a consequence of the more direct and open form of expression from an affectionate viewer.

Confusion Matrix: VADER

Accuracy 54.2%	Predicted Negative	Predicted Neutral	Predicted Positive
Negative	231	201	144
Neutral	18	33	31
Positive	41	23	278

To further demonstrate what has been indicated previously on a smaller sample of the predicted values, VADER classifies positives with a higher precision than negative, with a **true positive** rate of 81.2%.

Discussion

In this paper only one NLP method was actually implemented in order to provide an interpretable first result on crisis communication when performed by a comedian and YouTuber.

Although the video chosen is a particular case of "YouTube apology"[3], the data is highly available and contains the comments that are performed using typical responses: sarcasm and irony for negative messages, explicit messages for affection or sympathy, longer or unclear messages for neutral tones.

[3] On a semantic level, YouTuber Colleen Ballinger sings herself in the video in question: "*I'm not gonna take that route of admitting to lies and rumors that you made up for clout*": it is not an Apology video in the sense that the YouTuber posted it as such, but in the crowd's reaction to it.

Upon first results, VADER is not outperforming the expectations for a 10 year old lexicon-based model, although it correctly classified signs of affection or sympathy as they are usually more explicit in the comments.

XLNet shall be considered for future research, as its properties and training are commonly cited for its flexibility and accuracy.

In the aftermath of understanding the typical polarity of *apology videos*, in which there are usually three main categories of reaction, more comprehensive research shall be conducted by classifying the results based on the attributes of the three agents involved: the platform, the *apologizer*, the viewers. The platform can be set as a controlled setting, e.g., using YouTube for all comparisons of the comment sections, but more emphasis needs to be put on classifying the *apologizer* (their social and psychological attributes; their communication strategies and their performance) and the viewer(s)[4]: by using another NLP solution: topic identification. A possible model is **Latent Dirichlet Allocation (LDA)**, which allows for massive corpuses of text to be summarized into the few most common topics that are frequently mentioned, indicating more information on the tone of voice and how viewers perceived a certain medium.

Additional comments

No profiling, classification or identification of the user was performed in this research proposal; additional resources, code tutorials and materials utilized in this research proposal are referenced in the text or in the *References* section. Furthermore, no harm is meant to the reputation and intellectual or artistic property of the influencers and companies mentioned, as all the information reported is under the public domain.

The XLNet model was not used as extensively as VADER, given that the purpose of this research proposal is to explore and obtain one basis for future research.

[4] While we may more easily gather the information of a public figure such as influencers, it could be unethical and harmful to classify users based on their gender (as some research does, using their account information) and other personal information.

Bibliography

Patricia Bou-Franch, Nuria Lorenzo-Dus, Pilar Garcés-Conejos Blitvich, Social Interaction in YouTube Text-Based Polylogues: A Study of Coherence, *Journal of Computer-Mediated Communication*, Volume 17, Issue 4, 1 July 2012, Pages 501–521, https://doi.org/10.1111/j.1083-6101.2012.01579.x

Couldry, N., & van Dijck, J. (2015). Researching Social Media as if the Social Mattered. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115604174

Hall, J. A. (2018). "*When is social media use social interaction? Defining mediated social interaction"*. New Media & Society, *20*(1), 162-179. https://doi.org/10.1177/1461444816660782

Hutto, C.J. & Gilbert, E.E. (2014). "*VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media"* (ICWSM-14). Ann Arbor, MI. https://doi.org/10.1609/icwsm.v8i1.14550

Abd Kadir, S., Lokman, A. M., & Tsuchiya, T. (2016). Emotion and techniques of propaganda in YouTube videos. *Indian journal of science and technology*, *9*(S1). DOI: 10.17485/ijst/2016/v9iS1/106841

Kurtin, K. S., O'Brien, N., Roy, D., & Dam, L. (2018, May 31). *The development of Parasocial Interaction Relationships on YouTube*. The Journal of Social Media in Society. https://www.thejsms.org/index.php/JSMS/article/view/304

Park, S., Choi, JA. Comparing public responses to apologies: examining crisis communication strategies using network

analysis and topic modeling. *Qual Quant* **57**, 3603–3620 (2023). https://doi.org/10.1007/s11135-022-01488-5

J. K. Sandlin,M. L. Gracyalny. "*Seeking sincerity, finding forgiveness: YouTube apologies as image repair".* Public Relations Review, Volume 44, Issue 3, 2018, Pages 393-406, ISSN 0363-8111, https://doi.org/10.1016/j.pubrev.2018.04.007.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le (2019, June 19). "*XLNet: Generalized Autoregressive Pretraining for Language Understanding".* https://doi.org/10.48550/arXiv.1906.08237

References

Note: Regarding social media, the references are formatted using Columbia College guidelines (2024).

Github repository: "Strong Internet Language", by user Fluve (Davide Vandelli). https://github.com/FluveV/StrangersOnYoutube

YouTube API Services Terms of Service | Google for Developers

Ballinger C. [ColleenVlogs]. (2023, July 28). *hi* [Video]. YouTube. https://youtu.be/ceKMnyMYIMo?si=0vDXWnblh37W8PPO

Charles, J. *[Reupload from a different channel named* roblauren5270*] . (2019, May 11). James Charles weak apology video to Tati and James Westbrook.* [Video]. YouTube. https://youtu.be/PBV806YTXEg?si=QkDZ9qQMK0mpURPO

Columbia College (2024, January 5). APA Citation Guide (7th edition. APA Citation Guide (7th edition) : Social Media. https://columbiacollege-ca.libguides.com/apa/socialmedia

Ferragni, C. [@chiaraferragni]. (2023, December 18). *Sono sempre stata convinta che chi è più fortunato ha la responsabilità morale di fare del bene. Questi sono i valori…* [Video]. Instagram. https://www.instagram.com/reel/C0_wGLQIKPK/?utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA==

B. Fine, R. Fine. [*Reupload from a different channel* tacyppilF]. (2016, February 13). *Fine Bros Apology (REUPLOAD)* [Video]. YouTube. https://youtu.be/HoLSOba3_UE?si=iq1wXYli4Ghy9jT

Tsang, A. (2017, April 4) Nivea Pulls 'White Is Purity' Ad After Online Uproar. *The New York Times,* https://www.nytimes.com/2017/04/04/business/media/nivea-ad-online-uproar-racism.html

Jeffrey Lynn Steininger Jr. *also known as* Jeffrey Star. [*Reupload from a different channel* teaspillstation3231] (2020 July 19). *Jeffrey star APOLOGY VIDEO 2020* [Video]. YouTube. https://youtu.be/t5RALEy3b1M?si=WtU3S0lrm0o8SfqR