Detecting Offensive Language Using Linguistic Patterns: How Comparative Structures and Repetition Patterns Related to Offense?

Anonymous ACL submission

Abstract

This study collected and labeled 10,570 Mandarin comments from YouTube videos discussing controversial political issues, categorizing them into three types: "no offense," "attacking individual (direct)," and "attacking group (indirect)." The analysis reveals that incorporating features related to negative emphasis, comparison, and repetition significantly enhances model accuracy, which demonstrates that offensive speech often involves multiple linguistic strategies. This study contributes to the development of more accurate models for detecting offensive language, with particular relevance to Chinese linguistic contexts.

1 Introduction

With the increasing use of digital communication and social media, offensive language in public comments has become a rising concern for internet admins. According to Patton et al.(2016), social media plays an important role in intervening and preventing crisis related to community and gang violence. Risch & Krestel (2020) also points out that the internet misuse by spammers, haters, and trolls makes costly content moderation necessary. Offensive comments can contribute to the creation of toxic communities and impose a threat to users, therefore, developing an accurate model that detects offensive comments is a crucial task. Even though many online platforms already utilize models that rely mostly on keyword search, the efficiency and accuracy of the models has more to desire due to the subjective nature of offensive comments. Xu & Liu (2023) points out that while researchers can set rules to detect common offensive words, threatening phrases, or abusive speech. While these methods are relatively simple, they often require a lot of manual rule design and updating. Furthermore, despite many scholars improving the design of the mentioned models, little models are successful in detecting comments in Chinese. Xiao et al. (2024) states that Chinese, the second most spoken language, is a language that relies heavily on context to interpret the full meaning, that the same phrase may have distinct meanings under different circumstances or dialects; despite the increasing interest in detecting offensive language in Chinese, many studies fail to fully acknowledge the unique challenges posed by the language and culture, the prevalence of subversive expressions making it difficult for automatic systems to detect offensive language accurately. By focusing on the nuanced linguistic characteristics of Chinese offensive language, this study contributes to the broader understanding of how offensiveness is shaped by the use of linguistic patterns.

2 Studies on Offensive Language

2.1 Linguistic Perspective

Offensive speech can take on many forms. As a result, this study starts by identifying the common types of Chinese offensive speech and study examples.

The first type is sarcasm, which is a practice that makes fun of a certain person or group by stating the direct opposite of the quality the attacker wants to criticize. Xiao et al. (2024) defines sarcasm as "a sharp and often satirical or ironic utterance designed to cut or give pain". Despite sarcasm being a main form of offensive speech, Ibtissam & Azzeddine (2021) says that sarcasm is difficult to detect because it is implicit and characterized by the presence of positive words in a negative context, highlighting the problem with this specific type of offensive language. A potential solution is

suggested by Lin (2023) , by detecting phrases that show exaggeration or contrast that extends the meaning to an ungenuine extent, sarcasm could be detected.

The second type is impersonation, which is a practice where the user imitates the personality or identity of a target to slander the target.

Thirdly, mockery is a practice where the user also imitates the target, but unlike impersonation the imitation is usually exaggerated to make fun of the target. Frenda et al. (2023) describes sarcasm as a form of irony characterized by its biting nature and the intention to mock a victim.

Finally, Direct insult is a practice where the user use direct words or phrases to address the target, this type of offensiveness is mostly picked up by keyword models.

2.2 Machine Learning perspective

According to Lin (2023), sarcasm is often detected through ngrams, with the definition of sarcasm as "positive replies under negative comments" and the definition of irony as "positive utterance used to describe a negative situation". However, Lin (2023) also pointed out that Mandarin sarcasm detection studies mostly used only ngrams or features similar to ngrams, indicating there was a need to investigate whether features other than grams were needed.

Lin (2023) uses multiple cues for offensive comment detection:

The first type is exaggerating the negative connotations of the target. For example, "又", "簡直", "難怪".

The second type is emphasis on negative evaluations of the target. For example, "虧".

The third type is to emphasize the user's emotions to express that the target is wrong. For example, "咦", "啦", "捏".

The last type is to emphasize what the target should have done but didn't. For example, "就說", "不是說", "這才是".

3 Methodology

3.1 Data Collection and Labeling

This study collected YouTube comments from five videos by two famous YouTube channels. The YouTube channels post videos were chosen because they discuss controversial politic issues and have over one million subscribers. In total, 10570 comments are collected from the videos. The videos are the most recent on the channel during collection, and the videos were picked evenly from both channels. The comments' replies were also collected, with the comment they are replying to attached for training.

The data is collected through a scraper code that scraps down all comments in a video along with the comment's like count, user, time, and whom the comment replies to if it is a reply.

We recruited five college and high school students who are native Mandarin speakers as raters, they are rewarded cash equal to their efforts. The raters were given a two week work period to label the comments, and they agreed to let us use the data they labeled for this study.

The data was then split and distributed to six raters to label the comment. There is three types of labels: "no offend", "attacking individual (direct)", "attacking group (indirect)"

The followings are labels for data annotation (The instructions given to each rater):

- *label_1:* no offense

  definition: No offend to anyone

  example: 支持八炯 (Translation: "*Support Bajiong*")

  reason: Simply supports the YouTube channel

- *label_2:* attacking individual (direct)

  definition: Attacking the user the comment replies to, attacking the YouTube channel, or attacking to anything mentioned in video.

  example: 八炯你就是漢奸！我要獻上我最狠毒的詛咒來咒罵你😡你永遠是中國人！！全世界都是中國人，支持習主席再幹 50 年統治中國！習主席是世界的太陽，全世界都是中國不可分割的一部分🎉 (Translation: "*Bajiong, you are a traitor! I will offer my most vicious curse to curse you😡You will always be Chinese! ! The whole world is Chinese and supports President Xi to rule China for another 50 years! President Xi is the sun of the world, and the whole world is an integral part of China🎉*")

reason: The user directly attacks the channel, using words such as "traitor" and "curse".

- *label_3:* attacking group (indirect)

  definition: Attacking another group or individual.

  example: 共產黨就是世界毒瘤 (Translation: "*The Communist Party is a cancer in the world*")

  reason: Attacks the CCP party

Each comment is labeled by at least two raters, the final label of a comment will be "False" if both raters rate the comment as "no offend", otherwise the comment is labeled as "True". The kappa coefficient of our raters is 0.369, proving that offensiveness is a subjective aspect.

8553 comments are labeled as "True", while 2017 comments are labeled as "False".

3.2 Data Preprocessing

Comments under 3 characters or exceed 50 characters were deleted. Random "Offensive" comments are deleted for the number of both labels to be the same. After random shuffling the dataset the data is split into test and train sets. Using the TfidfVectorizer, the train set is fitted and transformed while the test set is transformed. For the baseline model a unigram logistic regression model from scikit-learn is used as it performed the best out of the machine learning models.

	"True" count	"False" count
Train set	462	511
Test set	1006	969

Table 1: Train and Test set Label Count

3.3 Incorporating Linguistic Features

Six features have been created in order to improve the baseline model, these features contain longer keywords that unigrams ignore and certain sentence patterns.

- *Feature 1:* Emphasis on Negative Aspect

  The feature is true if the comment contains 又 '*again*' , 難怪'*no wonder*', 當然'*for sure*', or contains the sentence structures [根本'*totally*'＋Negator].

- *Feature 2:* Comparison

  The feature will identify whether if the comment contains 比你'*than you*', 都比'*than*', 越來越'*more*', 不配'*not worthy of*' or contains the sentence structure [只有＋NP＋才].

- *Feature 3:* Downplay

  The feature will identify whether if the comment contains such as 反正'*just*', 不就是'*just*', 不過就'*just*',etc.

- *Feature 4:* Pointing out Impersonation

  The feature will identify whether if the comment contains such as 反串 '*pretending*', 了吧

- *Feature 5:* Emojis and Direct insults

  The feature will identify whether if the comment contains such as 笑'*laugh*', 可憐 '*pathetic*', 🤣, 😂, 🙃.

- *Feature 6:* Repetition

  The feature will identify the utterances that repeats a phrase or word for 6 times or more.

4 Results

4.1 Model Training Results

Model	min_df	Features	Accuracy
1+2+6	2	4412	0.732
Baseline	0	7301	0.731
Baseline	2	4406	0.724

Table 2: Results of Model Training

According to Tomaž & Kacijan (2020) repetition in a range of forms can contribute to building a malicious atmosphere in interrogation or torture scenes frequently start in a seemingly mild tone and progress to openly menacing and hostile

incidents, revealing the potential of repetition as an indicator of aggressive language.

According to Lin (2023), some emphasis phrases exaggerate or bring attention to a certain negative connotation, which potentially leads to sarcasm and therefore offensiveness. Also according to Lin (2023), contrasting phrases stretches the meaning of the comment to ungenuine directions. From the model training results, the combination of the features 1, 2, and 6 prove to exceed the performance of the baseline model. The modified model with the features 1, 2, and 6 significantly decreased the amount of features by roughly three thousand, while increasing the accuracy by roughly 0.1%. Therefore, this result suggest that offensive speech or language often feature multiple types of offensive speech, specifically the combine use of Emphasis on negative aspect, Comparison, and Repetition.

4.2 Implications From Model Training

Offensive language and its relation to comparative structures and repetition are topics of interest in the fields of linguistics and communication studies, particularly in examining how language functions in creating impact, asserting power, or maintaining social interactions. First of all, offensive language often involves comparisons, either implicitly or explicitly, to degrade or elevate individuals, groups, or objects. Comparative structures serve as linguistic tools to contrast different entities, often in terms of superiority or inferiority. When included within offensive or derogatory discourse, these comparisons amplify the offense by establishing a hierarchy or value judgment. For instance, Culpeper (1996) suggests that offensive language operates through impoliteness strategies, where comparisons can serve as a tool to threaten an individual's face (self-image). Comparisons often enhance the offensiveness by highlighting inferiority (e.g., "You are worse than a dog"). Studies of impoliteness show that comparative structures can reinforce the negative evaluation by juxtaposing a target with a disparaging standard (Bousfield & Locher, 2008). Repetition in offensive language can amplify the negative emotional impact and signal insistence, making the offensive expression more noticeable. Linguistically, repetition can serve as a rhetorical device that not only strengthens the derogatory intent but also deepens the offensive message into discourse. This tactic is particularly utilized in insults, bullying, and hate speech, where repetition emphasizes hostility and dominance. Tannen (2007) highlights that repetition in discourse can serve multiple functions, including intensification of meaning, which is often observed in confrontational or offensive interactions. Offensive language involving repeated insults or slurs can act as a form of "verbal aggression," where the repetition contributes to an ongoing power dynamic (Jay, 1992).

5 Conclusion

This study addresses the complexities of detecting offensive language in Chinese digital comments, with a focus on developing an improved detection model that exceeds the limitations of keyword-based approaches. Through the analysis of 10,570 YouTube comments on politically charged videos, we identified six critical linguistic features that enhanced model performance, especially those related to negative emphasis, comparison, and repetition. The integration of these features into the model demonstrated a noticeable improvement in accuracy, highlighting the multifaceted nature of offensive speech. Our findings suggest that offensive language frequently combines multiple offensive strategies, such as sarcasm, impersonation, and direct insults, making it more challenging to detect through traditional methods.

By focusing on the nuanced characteristics of Chinese offensive language, this study contributes to the broader understanding of how language functions in creating toxic online environments. Moreover, our model highlights the importance of contextualizing offensive language detection models to account for the unique linguistic features of the language being analyzed. Although the increase in accuracy may seem marginal (0.1%), the reduction in the number of features used in the enhanced model suggests potential scalability and efficiency in real-world applications. As offensive language continues to evolve, especially in multilingual and multicultural contexts, future research must consider further refinement of models to account for cross-linguistic subtleties and explore more sophisticated machine learning techniques. This study represents a step forward in developing more effective tools for moderating online discourse and fostering healthier digital communities.

6 Limitations

This study has several limitations. The lack of raters resulted in each comment only having two ratings, which is not ideal in interpreting subjective language. The collection of data is also limited in YouTube videos specifically about controversial politics, setting the risk of the model to be only able to accurately detect offensiveness in comments surrounding politics or news. By only using YouTube as the platform also might limit the scope of this study.

References

Derek Bousfield, and Miriam A. Locher, eds. 2008. *Impoliteness in language: Studies on its interplay with power in theory and practice* (Vol. 21). Walter de Gruyter.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics*, 25(3), 349-367.

Simona Frenda, Viviana Patti, and Paolo Rosso. 2023. When sarcasm hurts: Irony-aware models for abusive language detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 34-47). Cham: Springer Nature Switzerland.

Timothy Jay. 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets.* John Benjamins Publishing Company.

Jung Lin. 2023. *Sarcasm Detection in Mandarin Online Discourse* [Master's Thesis, National Taiwan Normal University]. Department of English, College of Liberal Arts.

Onic Tomaž, and Nastja P. Kacijan. 2020. Repetition as a means of verbal and psychological violence in interrogation scenes from contemporary drama. *Ars & Humanitas*, 14(1), 13-26.

Desmond U. Patton, Robert D. Eschmann, Caitlin Elsaesser, and Eddie Bocanegra. 2016. Sticks, stones and Facebook accounts: What violence outreach workers know about social media and urban-based gang violence in Chicago. *Computers in Human Behavior*, 65, 591-600.

Julian Risch, and Ralf Krestel. 2020. *Toxic comment detection in online discussions.* In Deep learning-based approaches for sentiment analysis (pp. 85-109).

Deborah Tannen. 2007. *Talking voices: Repetition, dialogue, and imagery in conversational discourse* (Vol. 26). Cambridge University Press.

Touahri Ibtissam, and Mazroui Azzeddine. 2021. Enhancement of a multi-dialectal sentiment analysis system by the detection of the implied sarcastic features. *Knowledge-Based Systems*, 227, 107232.

Yunze Xiao, Houda Bouamor, and Wajdi Zaghouani. 2024. Chinese Offensive Language Detection: Current Status and Future Directions. arXiv preprint arXiv:2403.18314.

Meijia Xu and Shuxian Liu. 2023. RB_BG_MHA: A RoBERTa-Based Model with Bi-GRU and Multi-Head Attention for Chinese Offensive Language Detection in Social Media. *Applied Sciences*, 13(19), 11000.