

Feature Name	Description
USERNAME FEATURES	
bot_str_in_name	1 if 'bot' appears in the username, 0 otherwise.
bot_str_at_beggining_in_name	1 if the username starts with 'bot', 0 otherwise.
anti_str_in_name	1 if 'anti' appears in the username, 0 otherwise.
utility_str_in_name	1 if keywords like 'reply', 'remind', 'link', 'repost', 'video', or 'image' appear in the username, 0 otherwise.
robot_in_name	1 if 'robot' appears in the username, 0 otherwise.
platform_str_in_name	1 if platform names (e.g., 'youtube', 'twitter') appear in the username, 0 otherwise.
bot_word_in_name	1 if 'bot' is identified as a distinct word segment in the username, 0 otherwise.
bot_word_capitalized_in_name	1 if the 'bot' segment in the username is capitalized (e.g., 'Bot'), 0 otherwise.
bot_word_uppercase_in_name	1 if the 'bot' segment in the username is uppercase (e.g., 'BOT'), 0 otherwise.
bot_word_at_beggining_in_name	1 if 'bot' is the first segment of the username, 0 otherwise.
bot_word_at_end_in_name	1 if 'bot' is the last segment of the username, 0 otherwise.
capitalized_name	1 if the username starts with an uppercase letter but is not camelCase, 0 otherwise.
uppercase_name	1 if the entire username consists of uppercase letters, 0 otherwise.
camel_case_name	1 if the username follows camelCase formatting, 0 otherwise.
#pl_words_in_name	Count of segments in the username identified as Polish words.
#eng_words_in_name	Count of segments in the username identified as English words.
#words_in_name	Total count of distinct word segments identified in the username.
words_in_name_char_avg	Average length of word segments in the username.
words_in_name_char_max	Length of the longest word segment in the username.
#chars_in_name	Total number of characters in the username.
number_in_name	1 if the username contains any digit, 0 otherwise.
number_at_end_in_name	1 if the username ends with a digit, 0 otherwise.
#special_chars_in_name	Count of non-alphanumeric characters in the username.
#underscores_in_name	Count of underscore characters in the username.
#hyphens_in_name	Count of hyphen characters in the username.
#digits_in_name	Count of digit characters in the username.
alpha_chars_density_in_name	Ratio of alphabetic characters to total characters.
special_chars_density_in_name	Ratio of special characters to total characters.
digits_density_in_name	Ratio of digit characters to total characters.
b_word_not_bot_in_name	1 if any segment starts with 'b' but is not 'bot', 0 otherwise.
username_char_entropy	Shannon entropy of the distribution of characters in the username.
username_char_uniformity	Character entropy normalized by the log of the number of unique characters.
CROSS LEVEL FEATURES	
#cross_lvl_words_unique	Count of unique username words that also appear in the comment text.
#cross_lvl_words_total	Total occurrences of username words found in the comment text.
polish_username_words_in_comments	Count of username words identified as Polish found in the comment.
english_username_words_in_comments	Count of username words identified as English found in the comment.
TEXT FEATURES	
type_token_ratio	Ratio of unique words to the total number of words.
top_10_words_ratio	Proportion of the text composed of the 10 most frequent words.
top_5_words_ratio	Proportion of the text composed of the 5 most frequent words.
word_entropy	Shannon entropy of the word frequency distribution.
word_uniformity	Word entropy normalized by the log of the number of unique words.
character_entropy	Shannon entropy of the character frequency distribution.
character_uniformity	Character entropy normalized by the log of the number of unique characters.
special_chars_entropy	Shannon entropy of special (non-alphanumeric) characters.

Table 4 – continued from previous page

Feature Name	Description
special_chars_uniformity	Special character entropy normalized by the log of unique special characters.
punctuation_entropy	Shannon entropy of punctuation characters.
punctuation_uniformity	Punctuation entropy normalized by the log of unique punctuation characters.
non_alpha_entropy	Shannon entropy of non-alphabetic characters.
non_alpha_uniformity	Non-alphabetic entropy normalized by the log of unique non-alphabetic characters.
whitespace_entropy	Shannon entropy of whitespace characters.
whitespace_uniformity	Whitespace entropy normalized by the log of unique whitespace characters.
#whitespace	Total count of whitespace characters.
whitespace_density	Ratio of whitespace characters to total characters.
#brackets	Total count of bracket pairs (round, square, curly, angle).
#round_brackets	Count of round bracket pairs '()'.
#square_brackets	Count of square bracket pairs '[]'.
#curly_brackets	Count of curly bracket pairs '{}'
#angle_brackets	Count of angle bracket pairs '<>'.
#nested_brackets	Count of brackets nested within other brackets (depth > 1).
chars_in_brackets_density	Ratio of characters inside brackets to lines of total characters.
words_in_brackets_density	Ratio of words inside brackets to total words.
#uppercase_words	Count of words composed entirely of uppercase letters.
uppercase_words_density	Ratio of uppercase words to total words.
#capitalized_words	Count of words starting with an uppercase letter.
capitalized_words_density	Ratio of capitalized words to total words.
word_length_avg	Average number of characters per word.
word_length_max	Length of the longest word.
word_length_std	Standard deviation of word lengths.
#urls	Count of URLs (http/https/www).
#markdown_links	Count of Markdown-style links.
starts_with_markdown_link	1 if the comment starts with a Markdown link, 0 otherwise.
#m_dashes	Count of m-dashes or double hyphens.
#elipses	Count of ellipses (...).
#hyphens	Count of hyphens.
#forward_slashes	Count of forward slashes.
#backward_slashes	Count of backward slashes.
#exclamation_marks	Count of exclamation marks.
max_consecutive_exclamation_marks	Length of the longest sequence of consecutive exclamation marks.
#question_marks	Count of question marks.
max_consecutive_question_marks	Length of the longest sequence of consecutive question marks.
#carets	Count of carets (^).
max_consecutive_carets	Length of the longest sequence of consecutive carets.
#rightwards_arrows	Count of rightwards arrows (>).
max_consecutive_rightwards_arrows	Length of the longest sequence of consecutive rightwards arrows.
#asterisks	Count of asterisks.
max_consecutive_asterisks	Length of the longest sequence of consecutive asterisks.
max_consecutive_punct_and_special	Length of the longest sequence of consecutive non-alphanumeric characters.
#quotes	Count of quotation marks (single or double).
#xd	Count of 'xd' string variants (e.g., 'xd', 'XDD').
xd_words_density	Ratio of 'xd' tokens to total words.
xd_non_alphanum_density	Ratio of 'xd' tokens to the count of non-alphanumeric characters.

Table 4 – continued from previous page

Feature Name	Description
#emojis	Count of emoji characters.
emoji_density	Ratio of emojis to total words.
#robot_emoji	Count of the robot emoji (U+1F916)
#emoticons	Count of text-based emoticons (e.g., ':)', ':-P').
emoticons_density	Ratio of emoticons to total words.
#slang_abbr_informal	Count of words identified as slang or informal abbreviations.
#sentences	Count of sentences.
sentences_capitalized_density	Proportion of sentences that start with an uppercase letter.
sentence_word_count_min	Number of words in the shortest sentence.
sentence_word_count_max	Number of words in the longest sentence.
sentence_word_count_avg	Average number of words per sentence.
sentence_word_count_std	Standard deviation of the number of words per sentence.
comment_ends_in_punctuation	1 if the comment ends with punctuation (!, ?, .), 0 otherwise.
#words_diacritics	Count of words containing Polish diacritics.
#words_potential_diacritics	Count of words found in a dictionary of de-diacritized forms (potential missing diacritics).
sarcasm_s_str	1 if the sarcasm indicator '/s' is found, 0 otherwise.
utility_str	1 if utility keywords (e.g., 'reply', 'save', 'mirror') are present, 0 otherwise.
priv_or_pv_word	1 if 'priv' or 'pv' is found, 0 otherwise.
platform_str	1 if platform names are found, 0 otherwise.
i_am_a_bot_str	1 if 'i am a bot' is found, 0 otherwise.
im_a_bot_str	1 if 'i'm a bot' is found, 0 otherwise.
m_a_bot_str	1 if 'm a bot' is found, 0 otherwise.
m_bot_any_str	1 if any 'm bot' variant is found, 0 otherwise.
jestem_bot_str	1 if 'jestem bot' (Polish) is found, 0 otherwise.
beep_boop_str	1 if 'beep boop' is found, 0 otherwise.
by_a_bot_str	1 if 'by a bot' is found, 0 otherwise.
i_am_a_bot_str_in_brackets	1 if 'i am a bot' appears inside brackets, 0 otherwise.
im_a_bot_str_in_brackets	1 if 'i'm a bot' appears inside brackets, 0 otherwise.
m_bot_str_in_brackets	1 if 'm bot' variant appears inside brackets, 0 otherwise.
beep_boop_str_in_brackets	1 if 'beep boop' appears inside brackets, 0 otherwise.
by_a_bot_in_brackets	1 if 'by a bot' appears inside brackets, 0 otherwise.
account_banned	1 if account status is 'banned', 0 otherwise.
#newlines	Count of newline characters.
newlines_sentences_ratio	Ratio of newlines to the number of sentences.
#unique_5grams	Count of unique 5-grams.
#unique_4grams	Count of unique 4-grams.
#unique_3grams	Count of unique 3-grams.
#unique_2grams	Count of unique 2-grams.
repetition_ratio_5grams	Ratio of repeated 5-grams to total 5-grams.
repetition_ratio_4grams	Ratio of repeated 4-grams to total 4-grams.
repetition_ratio_3grams	Ratio of repeated 3-grams to total 3-grams.
repetition_ratio_2grams	Ratio of repeated 2-grams to total 2-grams.
#polish_words	Count of words identified as Polish.
polish_words_density	Ratio of Polish words to total words.
#english_words	Count of words identified as English.
english_words_density	Ratio of English words to total words.
TEMPORAL FEATURES	

Table 4 – continued from previous page

Feature Name	Description
hour_of_posting	Hour of the day the comment was posted (0-23).
day_of_week	Day of the week (1=Monday, 7=Sunday).
is_weekend	1 if the comment was posted on a weekend (Saturday or Sunday), 0 otherwise.