Mapping and modelling word-final *n*-deletion in Dutch using Twitter data

by

Tommy Pieterse
s1010521

MA degree programme in
Linguistics and Communication Sciences (research)

Nijmegen, 16 December 2022

Supervisor: Dr. B.J.M. (Hans) van Halteren
Assessor: Prof. Dr. R.W.N.M. (Roeland) van Hout

Radboud University

Table of contents

Abstract

The pronunciation of word-final *-en* in Dutch constitutes a notable discrepancy between spoken and written Dutch for many speakers. Whether speakers pronounce *-n* at the end of the word depends on their region of origin, among a wide array of other factors. The rise of Twitter as a sociolinguistic data source provides a new opportunity to study the geographic distribution of word-final *n*-deletion and other non-geographic factors that influence its prevalence. The main research question guiding this study is: what can the examination of Twitter data tell us about the degree to which the occurrence of word-final *n*-deletion is distributed across the Netherlands and Flanders, and to what degree do internal and external linguistic factors influence its prevalence? The secondary research question was: to what degree is the use of Twitter data useful in mapping individual phonological features, especially in terms of the quality of the results? These questions were answered by automatically searching a large corpus of tweets in Dutch and submitting the resulting data to logistic regression and random forest classifier models. While we hypothesized that the resulting maps and results pertaining to the non-geographic features would mirror those relating to word-final *n*-deletion in spoken language, we instead found evidence that word-final *n*-deletion as used on Twitter constitutes a separate phenomenon. Therefore, the use of Twitter data also did not prove fruitful in the study of phonetic features per se—however, it does open the door to further research on this newly discovered form of word-final *n*-deletion.

1. Introduction

There are many notable discrepancies between written and spoken language. This is the case for most if not all codified languages today. Dutch is no exception. One of the main sites of discrepancy between spoken and written Dutch, as any new speaker will discover soon enough, is written word-final *-en*. This suffix has numerous functions, which includes signaling plural nouns, infinitives, and plural finite verbs. While its spelling might lead one to believe it should be pronounced like /ən/, it is most often realized as /ə/. To linguists, this phenomenon is known as word-final *n*-deletion. This study focuses on the occurrence of *n*-deletion[1] in written language sourced from the social media platform Twitter. Specifically, it is concerned with attempting to map its geographical spread using statistical model predictions as a baseline. This section deals with the investigation of scholarly literature on the background of word-final *n*-deletion and on language use on Twitter. The first subsection of this introduction will focus on exploring word-final *n*-deletion by going over existing literature detailing its history and the second subsection will explore the factors that are reported to influence its occurrence. Following this, Twitter as a new linguistic data source will be examined for its potential in the study of *n*-deletion. This section will conclude with an overview of the findings derived from the literature, which factors influencing *n*-deletion were selected to be included in the study, the research questions, and their corresponding hypotheses. In the section following this one, the methodology of the study will be laid out. This includes how the corpus of tweets was composed and how occurrences of *n*-deletion and *n*-retention were extracted from it, as well as the modeling and mapping methods applied to those occurrences. The section after that will display the findings that resulted from the reported analysis. The discussion that follows delves into the meaning of the results in relation to word-final *n*-deletion, the research questions, and hypotheses, followed by a discussion of the study's limitations and suggestions for future research. The study is concluded with a summary of the main findings and their significance in terms of *n*-deletion and the study of language on social media in general. The appendices attached at the end of this study contain additional information—Appendix 1 contains plots of the linear regression odds ratios generated for this study, Appendix 2 contains more in-depth results per model factor, including histograms, and Appendix 3 contains a document confirming approval of this study by Radboud University's humanities ethics assessment committee (EACH).

1.1 The origins of word-final n-deletion

While the exact period in which word-final *n*-deletion first arose in Dutch currently remains unknown, its first mention dates back to the 17th century. In his 1625 work on Dutch grammar, grammarian Christiaen van Heule (1625/1953) mentions that speakers in Holland (an area mostly coterminous with the modern-day provinces of North and South Holland) leave out the final /n/ when pronouncing words like *huyze*, *stede*, *lande*, *loope*, *blijve* and *valle* instead of pronouncing them as *huyzen*, *steden*, *landen*, *loopen*, *blijven* and *vallen* like other speakers in the Dutch language area (Van Heule, 1625/1953, p. 91). Important to note is that Van Heule (1625/1953) gives examples of word-final *n*-deletion in both nouns and verbs, indicating that it already constituted a phenomenon across multiple parts of speech by this time. That indicates it was certainly not limited to only certain words or phrases. Another fact to be noted is that Van Heule's (1625/1953) opinion on the occurrence of word-final *n*-deletion in speech is decisively negative: he considers the use of words with their word-final /n/s deleted to be in conflict with the nature of the Dutch language (p. 91). This demonstrates that word-final *n*-deletion likely finds its origins in speakers starting to drop word-final /n/ where they had previously not. A contemporary of Van Heule (1625/1953), Petrus Leupenius (1653/1958), in his discussion of verb conjugations, too, mentions that some have taken up the "bad habit" of

[1] The term "*n*-deletion" refers to *word-final n*-deletion specifically unless explicitly stated otherwise.

leaving the /n/ at the end of verbs ending on *en* unpronounced (p. 43). This confirms the presence of word-final *n*-deletion in verbs in some speakers and reinforces the assertion that word-final *n*-deletion was looked down upon at the time—or, more precisely, looked down upon by the scholarly community.

Importantly, cases of word-final *n*-deletion or mentions thereof in this time period are not exclusive to the works of grammarians. Indeed, words normally ending on *-en* with dropped word-final *n*s are also found in letters written during the 17th and 18th centuries by persons from the regions of Holland and Utrecht: Van der Wal et al. (2012), after having taken it upon themselves to examine the writing of a number of Dutch women from different social classes, report that the letters of their upper-class author show no word-final *n*-deletion. They link this to the orthographic practices of authors of printed texts at the time, which also show no word-final *n*-deletion. This, combined with Van Heule's (1625/1953) clear disapproval of word-final *n*-deletion, all but certainly indicates it was not part of the standard language (to the degree to which a standard language can already be said to have existed during this time) in writing and speech. The letters of a lower-middle-class author in their set, however, systematically display written word-final *n*-deletion. Notably, the author's application of word-final *n*-deletion appears to be highly regular and largely rule-based. She deletes 100% of the time in the case of plural nouns and finite verbs, 84% of the time in the case of infinitives, and 0% of the time in the case of past participles and singular nouns ending on *en*. While we do not know anything about this author's spoken language directly, considering the non-standard status of word-final *n*-deletion in both spoken and written language at the time, the fact that word-final *n*-deletion appears in her writing likely indicates it was present in her spoken language as well. The *n*-deletion patterns (or lack thereof) in the writing of these women indicate that the occurrence of word-final *n*-deletion during the 17th and 18th centuries was likely already highly regular in terms of rules and that it carried non-standard connotations. It should be noted, however, that Van der Wal et al.'s (2012) dataset is considerably limited in terms of the number of individual authors. This means that more thorough research is needed to paint a conclusive picture of the historical presence of word-final *n*-deletion in Dutch.

1.2 Factors influencing the occurrence of word-final n-deletion

To understand word-final *n*-deletion, knowledge of merely its historical context is not sufficient. It is equally important to examine exactly what factors possibly exert influence over its manifestation in a linguistic sense. Fortunately, word-final *n*-deletion has received quite some attention from the scholarly community. This attention manifests itself largely in the form of quantitative linguistic research projects, starting with a study by Ollevier (1959). While the original text of this unpublished paper has been lost to time, some of its findings and their implications can be retrieved from a small discussion article written by Pauwels (1969). In an effort to ascertain how Flemish speakers of Dutch render *-en*, Ollevier (1959) monitors radio broadcasts and Flemish university professors, leading to a sample of 335 Flemish standard Dutch speakers. Ollevier (1959) constructs his study to take into account differences between word types (nouns vs. verbs, strong past participles vs. other verbs, and stems vs. composite forms) and the type of sound following word-final *-en* (vowel, consonant, or a pause), but Pauwels (1969) does not report on any findings specific to these categories. Ollevier's (1959) main finding is that his pool of subjects seems to be split roughly down the middle, with 85 speakers deleting /n/ in over 50% of opportunities and 103 speakers maintaining /n/ in over 50% of opportunities. Furthermore, around half of the participants had a deletion or realization rate of over 90%, indicating that many speakers have a clear preference. In a limited number of cases, Ollevier (1959) was able to explore the context of certain recordings, during which he found limited support that word-final *n*-deletion rates are affected by the genre in which the words are spoken, with the reading of texts within a broadcast context and poems leading to

lower rates of word-final *n*-deletion. According to Pauwels (1969), Ollevier (1959) concludes that word-final *n*-deletion has entered or is entering standard speech in Flanders, although it seems to be holding out in some more, as Pauwels (1969) puts it, "ceremonious" domains (p. 218).

A subsequent quantitative study on word-final *n*-deletion was conducted by Koefoed (1979), who studies a total of 120 minutes of Dutch broadcast radio recordings and supplements these with recordings of 10 participants who were instructed to read and repeat texts out loud. For every individual speaker, Koefoed (1979) computes an *n*-realization score that corresponds to the proportion of times each speaker realizes /n/ out of all total possible instances that a speaker could have realized /n/. He finds that the main factor in his data that influences the realization of /n/ is style, with /n/ being realized more often in cases where speakers are reading texts aloud, and /n/ being deleted more often in cases where speakers engage in spontaneous conversation. Notably, Koefoed (1979), too, notes that there is a large degree of inter-speaker variation, using the example of two female newsreaders who differ considerably in the number of times they realize /n/ to illustrate his point. Koefoed (1979) hypothesizes this might be due to a difference in dialect backgrounds between the speakers yet recognizes that investigating dialect background as a factor is not possible due to the limitations of his dataset.

After the publication of the study by Koefoed (1979), studies with more complex designs begin to emerge. The first of these is a study by Van Oss and Gussenhoven (1984), who focus their attention on Dutch TV newsreaders and their *n*-deletion behavior, specifically with regard to nouns. They take into account a relatively large number of features. Firstly, they differentiate between singular (monomorphemic) nouns (e.g., *deken* ("blanket")) and plural (polymorphemic) nouns (e.g., *daken* ("roofs")). They report being unable to find a significant difference between the two categories in the vast majority of speakers. Secondly, Van Oss and Gussenhoven (1984) look at the type of sound that follows word-final *en*. They report that word-final *n*-realization is most common immediately before a vowel, more so than immediately before a consonant or a pause. Lastly, they incorporate the age of the speakers into their design but only do so for their sample of newsreaders. They report a correlation between age and word-final *n*-deletion, with younger speakers deleting more often than older speakers. They interpret this as a sign that word-final *n*-deletion is in the process of spreading through standard Dutch. Important to note is that, when examining the data on a per-speaker basis, they propose two different types of speakers: "deleters" and "inserters." The speech of the deleters is marked by a high degree of word-final *n*-deletion, with speakers also deleting more /n/s in plural nouns than in singular nouns. The speech of the inserters is marked by a generally high degree of word-final *n*-realization, with /n/s being realized more often in plural nouns than in singular nouns.

As part of her doctoral thesis on the speech of "locally prominent" (i.e., upper-middle-class) speakers of standard Dutch in the Dutch towns of Middelburg, Roermond, and Zutphen, Voortman (1994) devotes a small section of her study to word-final *n*-deletion. In addition to considering the speakers' towns of origin, Voortman (1994) examines both formal and informal conversations. Concerning formal contexts, she reports not finding any significant differences between the three towns: speakers from all towns almost always realize their /n/s, with the main percentage of *n*-deletion out of all possible instances being lower than 10%. Only in the case of the speakers from Zutphen was Voortman able to record informal speech. Here, /n/s are deleted more often than they are realized (77.4% of possible /n/s were deleted). This once again points to a difference in formality as an important factor in the deletion of word-final *n*. Voortman (1994) considers the formality effect a consequence of the speakers recognizing differences in style and shifting their speech accordingly, although she does not completely rule out the possibility that attention to speech might also have had an impact on her participants' rate of deletion.

Another study of word-final *n*-deletion that forms part of a doctoral dissertation is that of Van de Velde (1996), who examines the spontaneous speech of radio reporters in both the Netherlands and Flanders. Unlike the authors of previous studies, Van de Velde (1996) attempts to limit the variation in ages of the reporters at the time of recording as much as possible, while only varying the dates of the recordings themselves. This effectively eliminates the possibility of age grading as a confounding factor in the investigation of the possibility of ongoing language change. Interestingly, contrary to previous studies, he finds no evidence that word-final *n*-deletion is spreading in standard Dutch. What does match with the findings of previous studies is that speakers generally realize /n/ most often immediately before a vowel. Van de Velde (1996) also examines possible differences between the Netherlands and Flanders. While both Dutch and Flemish speakers exhibit relatively high rates of deletion, there is more inter-speaker variation in the Flemish sample, with the groups of deleters and inserters showing up clearly within that set of speakers. He also finds an effect of the word category (in this case infinitive verb, finite verb, or plural noun) within the set of Dutch speakers and not among the Flemish speakers, but he simultaneously acknowledges that a more detailed study is required to make further claims regarding this factor.

In a study that followed relatively soon after Van de Velde (1996), Van de Velde and Van Hout (1998) attempt to delve deeper into inter-speaker variation in *n*-deleting behavior. They also investigate the speech of standard Dutch radio speakers and their rate of word-final *n*-deletion. Instead of the two groups delineated by Van Oss and Gussenhoven (1984), Van de Velde and Van Hout (1998) differentiate four different types of speakers. The first of these is the "non-realisers." This group is characterized by total word-final *n*-deletion, implying that they have no underlying rule for pronouncing the /n/ in *-en* anywhere. The second of these is the "liaisoners," whose *n*-realization behavior resembles that of French *liaison*. They do delete /n/ before consonants and pauses, but they do generally exhibit some (anywhere from 13-67%) realization before vowels. For these speakers, *n*-realization is a post-lexical rule—i.e., it is not influenced by grammatical concerns like the morphological status of the word. The third group is a group named "deleters," which should not be confused with the group with the same name as defined by Koefoed (1979). These speakers exhibit *n*-realizing behavior in all right-hand contexts. This behavior is affected by the morphological type of the word, with monomorphemic nouns having their word-final /n/ deleted less often. Lastly, there is the group of "pausers," who realize /n/ most often before a pause, essentially transforming it into a discourse marker. The *n*-deletion rate of this group is also subject to a morphological effect. The study by Van de Velde and Van Hout (1998) symbolizes one of the first major successful attempts to tease out the different possible internal rule systems that determine word-final *n*-deletion in speakers of Dutch.

Van de Velde and Van Hout (2001) shift their attention to a different set of speakers. Instead of investigating radio broadcasts, they derive their speech data from a corpus of teachers from the Netherlands. This corpus is stratified by region, gender, and age. In addition to the demographic factors taken into account by the corpus itself, Van de Velde and Van Hout (2001) also incorporate word type into their study, differentiating between monomorphemic nouns, monomorphemic verbs, polymorphemic finite verbs, polymorphemic infinitive verbs, and spatial adjectives and prepositions. After subjecting the corpus to a thorough analysis, they uncover a regional difference: speakers in the North of the Netherlands realize /n/ more often than those in other parts of the country. They also report an age-gender interaction: young women seem to be deleting /n/ most often, while young men delete it least often. Age as a factor by itself was not significant, further bolstering the assertion that word-final *n*-deletion is not in a state of expansion. As for word type, word-final /n/s are realized most often in monomorphemic verbs, followed by monomorphemic nouns, polymorphemic finite verbs, spatial adjectives and prepositions, and finally polymorphemic infinitives.

After their 2001 study, Van de Velde and Van Hout (2003) expand their approach to include the right-hand context of the occurrences of *-en*, while also providing a more in-depth analysis of the data. An important finding is an interaction between country/region and gender: in the Netherlands, men realize /n/ more often than women, but the opposite is true for Flemish speakers. Van de Velde and Van Hout (2003) also explore more specific regional differences. They find that /n/-realization occurs mostly in the North of the Netherlands and in the provinces of West and East Flanders, which comprise the westernmost third of the region of Flanders. West and East Flanders also turn out to be home to the largest degree of inter-speaker variation, possibly due to hypercorrection. The so-called "suffix effect"—i.e., that word-final /n/ is realized more often in finite verbs than in non-finite verbs—only exists in the Netherlands. Findings further include that the effect of the morpheme status of a word (i.e., monomorphemic vs. polymorphemic) is significant in both the Netherlands and Flanders. Van de Velde and Van Hout (2003) also report a so-called "focus effect," which entails the phenomenon of speakers being more able to steer their language before pauses, which means that speakers are most likely to realize or delete /n/ before a pause, depending on which of those has their preference. They conclude their study by reaffirming their stance that there is no single, language-wide rule system behind the deletion of word-final /n/ and argue that out of all factors, right-hand context is likely to be the most significant.

An honorable mention should be made of a study on *n*-deletion conducted by Goeman (2001), which provides the most in-depth information on the geographical spread of *n*-deletion in the Netherlands. However, Goeman (2001) investigates not only word-final *n*-deletion but also word-internal *n*-deletion and leaves Flanders out of the scope of his study. Therefore, the maps he generates are limited in their use for the study of word-final *n*-deletion in the Dutch language area. Nevertheless, considering their great level of detail, these maps deserve some attention. What follows is a brief summary of the main characteristics of some of those maps. The map showing *n*-deletion before a pause shows a low degree of deletion in the northern and eastern parts of the Netherlands, medium rates in the Randstad area, and very low rates in Brabant and Limburg. The map detailing the deletion of /n/ immediately before vowels shows relatively low rates of deletion throughout the country (0–42% deletion), but the lowest rate (28–42%) can be found in the western Randstad area. The map showing *n*-deletion before consonants show the highest rate of deletion occurs in the extreme south of the province of Limburg and that the lowest rate of deletion occurs in the North of the Netherlands, with the rest of the country falling somewhere between the two areas. The remaining maps that display specific word types all generally show the lowest rate of deletion in the North of the Netherlands and the lowest rates in the South and West of the Netherlands. As stated before, however, as word-internal *n*-deletion is also included in these maps, the data these maps display cannot be used to determine where word-final *n*-deletion occurs in the country to a high level of detail. It does, however, provide a general image of where word-final *n*-deletion is likely to occur, namely most often in the West and South of the Netherlands and the least in the North and North-East.

To summarize, we can, based on the existing literature, distinguish a number of factors that are likely to influence word-final *n*-deleting behavior. The first of these is formality. Generally, speakers are more likely to realize word-final /n/ in formal settings and more likely to delete it in informal settings. Secondly, the right-hand context of *-en* seems to play an important role, with word-final /n/ being realized more often before vowels, although pauses allow for the speaker to express their preferred way of pronouncing *-en* more than other following contexts. Thirdly, the speaker's region of origin influences word-final *n*-deletion as well, with speakers from the North and North-East of the Netherlands and speakers from West and East Flanders leaving /n/ undeleted most often. While a speaker's age does not seem to be a significant factor by itself, it does exhibit interactions with gender and region, with (1) young

women deleting the most, and young men deleting the least, and (2) Dutch women deleting more than Dutch men, and Flemish men deleting more than Flemish women. The word type in terms of part-of-speech and morpheme status has an influence on the pronunciation of *-en*, with monomorphemic verbs ending on *-en* exhibiting the lowest degree of deletion, followed by monomorphemic nouns, polymorphemic finite verbs, spatial adjectives and prepositions, and, lastly, polymorphemic infinitives. All of these factors, however, are influenced by a speaker's underlying set of rules. Based on those rules, speakers can be divided into non-realizers, liaisoners, deleters, and pausers. While a great deal of information has already been uncovered by scholars, it is also apparent that a more in-depth study of one or more of these factors has the potential to shed more light on the processes that influence word-final *n*-deletion.

1.3 The potential of Twitter as a source of linguistic data

One possible method of increasing knowledge of word-final *n*-deletion is to venture into new data sources. One of those potential new data sources is Twitter. It might seem like an unlikely candidate, as word-final *n*-deletion is mostly limited to spoken Dutch. However, some studies indicate that it might prove more useful than one would first be led to expect. This is because some users transpose features of their (colloquial) spoken language into the writing of the tweets they author. Hilte (2019), in her doctoral dissertation on the social media language habits of teenagers, elaborates on this *orality principle*. Authors of social media posts adapt their writing in a number of different ways to increase its orality (i.e., bring it closer to spoken language). They might, for example, leave out certain letters of a word to make it seem closer to its pronunciation in spoken language, they might add letters to indicate stress, or they might choose to choose certain lexemes that are more appropriate to spoken language. She also reports on the effects of gender and age, with women using more online-specific nonstandard forms than men and younger adolescents using more variation in spelling to express themselves.

Over roughly the past decade, a number of linguistic studies specifically employing tweets as their principal source of data have sprung up that incorporate the possibility of aspects of spoken language carrying over onto Twitter. These studies indicate that the orality principle is also applicable to the written language used in tweets, and in a number of ways. The first of these is communication accommodation—i.e., the phenomenon that interlocutors tend to converge in terms of linguistic and other communicative aspects. A study by Danescu-Niculescu-Mizil et al. (2011) delves into whether communication accommodation also occurs on Twitter. They construct a probabilistic framework based on, *inter alia*, stylistic markers in tweets, which they employ to measure communication accommodation. They find that the language of tweeters in a conversation tends to converge just like it has been shown to in spoken contexts. They emphasize the fact that this is the first time that communication accommodation has been tested outside of small-scale experimental settings, which already highlights one of the important advantages of using tweets: researchers have access to a gigantic pool of language data, the size of which would be utterly unapproachable by most "classic" linguistic data sources.

Other studies that incorporate Twitter-sourced data into their design focus on sociophonetic variation. Eisenstein (2015) investigates a number of features associated with the dialect of African-American English in a corpus of 114 million tweets. Using logistic regression, he finds words that deviate from the orthographic standard and which approximate a more phonetic spelling (e.g., words that normally end on *-ing* being spelled with *-in*) are frequent and carry much of the social meaning of the non-standard or dialectal spoken forms they mirror. Notably, he reports that the rate at which words are spelled in a non-standard manner differs per word type (e.g., verbs vs. nouns). The phonological context of a word also seems to be a weaker factor when compared to its importance in spoken sociophonetic variation.

While Eisenstein's (2015) study is largely quantitative in nature, qualitative approaches to sociophonetic orthographic variation on Twitter have also appeared. One such study is that conducted by Ilbury (2019). He focuses on the use of features normally belonging to African-American English in the tweets of 10 gay British men. He argues that these men employ these features in order to construct what he calls a "Sassy Queen persona," which relies on drawing upon stereotypical characteristics associated with African-American women like fierceness and sassiness. This demonstrates that, much like with linguistic variables in spoken language, orthographic variation in tweets can also be used to construct identities.

The use of Twitter data has also proven useful in the study of style-shifting. Tatman (2015) directs her attention to the encoding of phonetic variants that carry specific social meaning in spoken language and their presence on Twitter. She specifically focuses on features common in white vernacular dialects of the US South and African-American English in her first set of tweets and features common to Scottish English in a second set of tweets. Based on the language use of one Scottish tweeter, whose tweets were examined in more detail, she determines that the use of socially meaningful phonetic spellings is sensitive to topic-based style-shifting, as is also the case for the spoken-language counterparts of the investigated spellings. It should be noted, however, that Tatman's (2015) study remains cursory in that it is limited by its relatively small sample size. While the total number of sampled tweets is not explicitly stated, inspecting the presented graphs points to a sample size of around a few hundred tweets. Therefore, the study does not yet make full use of tweets' potential in numbers. It does, however, pave the way for more large-scale studies focusing on style-shifting on Twitter.

One of those more-large scale studies is a study by Shoemark et al. (2017). They focus on investigating the occurrence of topic-based style-shifting in a corpus of around 30,000 Scottish Twitter users. This set consists of two different groups of users, with one group exhibiting a higher degree of use of Scottish-English. They incorporate around some 50 lexical variables into their study that represent standard English lexical items and their Scottish-English equivalents, such as *don't* vs. *dinny*. They use a Latent Dirichlet Allocation topic model to divide the dataset into topic categories that were then manually labeled. They employ a mixed-effects logistic regression model to process their data and find that the language use of both user groups is affected by the topic of a tweet. The groups did, however, differ in the degree to which they shifted, under which conditions they shifted, and toward which language they shifted.

As it turns out, the study of the lengthening of sounds has also employed tweets as a valuable source of data. Gray et al. (2020) use a Twitter corpus to examine whether spoken language carries over into the written language of tweets. Their corpus consists of roughly 10% of English tweets between 2008 and 2016, collected using Twitter's *Gardenhose* API. They specifically focus on "stretchable words," like "duuuuude" and "yeeeees." While their further investigation into these stretchable words lies outside of the scope of this paper, an important takeaway is that these stretchable words, too, resemble spoken language and are used in similar contexts.

1.4 Using Twitter data to map linguistic features and languages

Perhaps one of the most important factors indicating the potential of Twitter data to study word-final *n*-deletion does not lie in the text of the tweets themselves. Instead, it lies in the meta-data associated with certain tweets and profiles to link certain users and, by extension, their writing to certain locations. There are two main possible sources for this location data: the location that is part of a user's profile, which is self-provided and therefore rather error-prone, and GPS coordinates attached to individual tweets, which are more accurate but are also rarer. This section will examine linguistic studies that use these data sources to generate maps based on

certain linguistic features or languages. The studies are divided by what type of linguistic feature(s) they investigate: studies on lexical features will be discussed first, followed by studies delving into phonetic features, studies delving into morphosyntactic features, and, finally, studies that fall outside of these categories. It should be noted that these categories are sometimes rather vague, and the topics of some studies might bleed into other categories.

Gonçalves and Sánchez (2014) focus on using lexical alternations (e.g., *computadora* vs. *ordenador*) to map variation in the Spanish language. They collect some 50 million Spanish-language tweets, of which 750 thousand include geolocation metadata. They divide the Spanish-speaking world up in a grid pattern based on latitudinal and longitudinal degrees, with each square having sides of roughly 25 km by 25 km at the equator. They cluster the users in their dataset using principal component analysis and uncover two major superdialects, which both span both Latin America and Spain. The first superdialect is centered around urban regions and is characterized by a high rate of innovation, while the second superdialect is centered mostly around rural regions and is much more conservative in nature. They further split the second superdialect into three regional dialect areas, which center around Spain, northern Latin America, and southern Latin America. This study represents the first large-scale effort to use Twitter location data to map linguistic features, but it is certainly not the last.

Doyle (2014) directs his attention to the phrase *needs done* and so-called double modals (e.g., *might could*) in US English. While one could argue these double modals do not necessarily constitute lexical features, Doyle (2014) investigates them in much the same way one would in the case of lexical features by focusing on the occurrences of certain static phrases. He gathers tweets from the entire US and uses words like *the* and *I* in his search process to ensure the collected tweets form a solidly English base. An investigation into the prevalence of *needs done* reveals that its occurrence in the Twitter corpus lines up remarkably well with spoken language data. The same is true for the double modal data. Doyle (2014) attributes the success of these maps to his method: because of a lack of "negative data" (i.e., data on cases in which the phrases were not used), he uses what is known as Bayesian inversion in order to account for the overall distribution of the feature, not just the non-standard form.

Eisenstein et al. (2014) focus not just on the occurrence of certain lexical items but also investigate the diffusion of those items. Their corpus, consisting of 107 million tweets authored in the US by some 2.7 million unique users, is used to trace the spread of new lexemes like *ion* ("I don't"), *af* ("as fuck") and *ikr* ("I know, right?"). They construct a system—whose exact workings lie outside of the scope of this paper—that includes a latent vector autoregressive model and logistic regression. They uncover several important factors that influence the rate and the path of lexical diffusion. While geographical proximity and the population size of a user's town of residence have some influence, the most important factor is demographic similarity, especially race. Rather than some unified "netspeak" slowly arising through the internet, Eisenstein et al. (2014) argue that their data indicates that online language use mirrors the fault lines that exist between the different dialects of spoken US English.

As is the case for two of the three previously discussed studies, Jones (2015) takes the United States to be his main area of investigation. His chief purpose is to investigate internal variation in African-American English, a topic that had up until that point been relatively neglected by scholars. He takes into account around 30 common words originating from African-American English and non-standard spellings of certain words and phrases that possess a considerable degree of use among African-Americans. Examples of these items include *finna* ("going to"), *sumn* ("something"), and *yuon* ("you don't"). He attempts to collect at least 10 thousand tweets for each item, in which he mostly succeeds. Using only the GPS data attached to specific tweets to determine users' locations, he maps out the occurrences of these features and discovers that these maps exhibit apparent dialect regions within African-American English that differ in the degree to which spellings and lexical items are used. Importantly, these dialect

regions align with historical migration patterns. This indicates that, even though this study reveals "new" dialect regions, the existence of these regions is not likely to be spurious.

Donoso and Sánchez (2017), like Gonçalves and Sánchez (2014), use a grid pattern to map their Twitter data. In many ways, the former seems to be a sequel to the latter. They both focus on alternations of Spanish lexical items, and they both seek to cluster their data to tease out dialect areas. In total, their corpus comprises around 11 million geotagged Spanish tweets. Donoso and Sánchez (2017) differ from Gonçalves and Sánchez (2014) in that they focus merely on Spain and exclude Latin America from their data. They also differ in their clustering methods, using cosine similarity and Jensen-Shannon divergence to divide their sample into two major clusters, measured over all lexical variants. These clusters, again, seem to differ in that one consists mostly of urban users, and that one consists mostly of rural users.

If one were to shift their attention back to the United States once more, they would be directed to a study by Eisenstein (2017). He constructs a corpus of geotagged tweets that originate in the United States and range from the year 2009 to 2012, ending up with 114 million messages authored by 2.77 million individual users. What sets his approach apart from previous ones is that, instead of attempting to scour the data for the occurrence of pre-determined lexical variants, he sets out to explore the data without defining any lexical variables beforehand. He does this in the hopes of discovering dialect regions first, after which these regions can be explored for their distinguishing characteristics. This approach proves fruitful, and his process leads to dialect regions that in many ways seem to line up with dialect regions as defined by previous studies, although these maps are not a perfect match. A diachronic analysis also reveals that newer lexical items (including new abbreviations) like *ard* ("alright") and *ctfu* ("crack the fuck up") are rapidly spreading, which again indicates that Twitter data might also be useful for the study of lexical diffusion. He concludes by setting the exploration of areas outside of the US as a major priority, as these areas—especially those characterized by variation continua instead of discrete dialect regions—might also offer a considerable amount of useful information on lexical variation and diffusion.

Grieve et al. (2018), too, direct their attention to the United States. They present a highly detailed study of lexical diffusion using Twitter. Their corpus is expansive, containing 980 million tweets authored by 7 million unique users during 2013 and 2014. They focus on 54 emerging lexical items, like *mce* ("man crush everyday"), *notifs* ("notifications"), and *boolin* ("hanging out") and their orthographic variants. They map these lexical items' cumulative relative frequency at different dates. These maps reveal that new lexical items tend to appear in a small number of regional hubs, after which these items spread along generally consistent pathways of distribution. The results indicate that an area's cultural relevance is more important than population size when it comes to actuating lexical diffusion. Grieve et al. (2018) then conduct a multivariate spatial analysis, which leads them to distinguish five main regional patterns of lexical innovation. These regional patterns can be distinguished by which US macro-region they span.

Grieve et al. (2019), acting in accordance with Eisenstein's (2017) recommendation to venture outside of the US, focus mainly on lexical variation in the United Kingdom. They assemble a corpus consisting of 180 million geolocated tweets authored by 1.9 million unique users. By their own admission, they leave this corpus completely unfiltered, stating that they "believe that modifying the corpus to make it more likely to show regional patterns is a highly subjective process that necessarily results in a less representative corpus" (Grieve et al., 2019, p. 4). They scan this unfiltered corpus for lexical alternations (like *couch* vs. *sofa* vs. *settee*) and use the spatial correlation coefficient *L* to compare their maps to the BBC Voices dialect survey. The results of their comparison show that there is generally a large degree of alignment between the Twitter and BBC maps, but they also suspect that low-to-mid *L*-values still indicate a significant degree of alignment, leaving some things to be desired.

Clearly, the use of Twitter data to map out the spread of certain lexical variables is by now relatively well-established. The same cannot be said for phonological features. Out of all Twitter-based mapping studies, only one focuses partly on phonological variation. Van Halteren et al. (2018) focus on lexical and phonological dialect features of Limburgish, a regional Germanic (group of) dialects in the Netherlands. Using the pre-existing TwiNL corpus of Dutch tweets (Tjong Kim Sang & Van den Bosch, 2013), selecting tweets authored between 2011 and 2017, they end up with about 7200 users who exhibit some form of Limburgish in their tweets. Their knowledge-rich approach focuses on detecting 15 patterns that match the orthographic equivalent of Limburgish phonological forms and their standard Dutch counterparts, while their knowledge-poor approach focuses on trigrams. They smooth the resulting user scores over each user's 99 nearest neighbors, with each neighboring user's influence being proportional to the distance between the user and that neighbor. They also experiment with log-scale binning the scores, and the resulting maps of the individual phonological features line up considerably well with knowledge of the isoglosses of Limburgish that were established in previous studies. This study is especially relevant to this paper as it focuses on a phonological feature within the Dutch language area.

Besides studies on lexical and phonological features, two studies dedicate themselves to studying morphosyntactic features. The first of these is Stevenson (2016). He focuses specifically on the different word orders associated with ditransitive constructions in English in the United Kingdom. These different word orders can be categorized into three types, namely *PDAT* ("Send it to me," the most common variant), *GTD* ("Send me it"), and *TGD* ("Send it me"). He constructs a small corpus of 1416 geo-coded tweets originating from dates between November 2014 and March 2016. He maps the relative frequency of the constructions, and the resulting data matches existing dialect survey data decently well. He does, however, acknowledge the limited size of his dataset and emphasizes that his study is but a cursory exploration of the use of Twitter-based corpora and their relevance for the study of dialectal variation.

The second study that focuses on mapping morphosyntactic variation is that of Willis (2020), which focuses on morphosyntactic variation in Welsh. He constructs a dataset of tweets in Welsh, taking into consideration both tweet-specific GPS data and user-supplied locations. His justification for doing so is that one cannot afford to discard too much data when dealing with smaller datasets. In the first section of his paper, he specifically inspects variation in the second-person singular informal pronoun. In the second section, he focuses on auxiliary *bod* ("to be"). In both cases, he also takes syntactic context into consideration. He maps the data using KDE (kernel density estimation) smoothing, which calculates the mean frequency within a kernel centered on a given point. The resulting maps line up with existing knowledge of these linguistic variants as they are distributed in spoken Welsh. Willis (2020) does point out that the frequencies of certain variants do not always accurately resemble those found in spoken language surveys, but this does not take away from the large degree of alignment between his Twitter-based maps and dialect survey maps.

One study that is concerned with mapping linguistic data using Twitter does not limit itself to only one domain of features and instead incorporates lexical, phonological, *and* morphosyntactic features into its design. Ljubešić et al. (2018) construct a corpus of about 1.7 million Serbo-Croatian geocoded tweets by 17 thousand unique users. The purpose of their study is to attempt to investigate the boundaries between the different varieties of the multi-centric language Serbo-Croatian (including Serbian, Croatian, Montenegrin, and Bosnian). They consider a broad array of lexical, phonological, and morphosyntactic features that factor into a speaker's position on the Serbo-Croatian continuum. They draw the geographical boundaries between the features they investigate, partly with the help of KDE smoothing. Their results show that Serbian and Croatian, as backed up by existing research, form the two

extremes of the Serbo-Croatian continuum, with other areas leaning more one way or the other depending on what feature is inspected.

An honorable mention goes to the study by Abitbol et al. (2018), which attempts to link French Twitter data to socioeconomic status, age, location, time, and social networks (in the sociolinguistic sense of the word). Firstly, they map French socioeconomic survey data, after which they use Twitter data to map (a) a user's rate of correct use of the French negation construction, (b) their rate of correct plural forms, and (c) their vocabulary size. They do this on the level of individual French departments. They compare their data to existing socioeconomic data and report that a higher degree of standard language use is correlated with a higher socioeconomic status. Furthermore, the South of France seems to be using more standard language than the North.

Several important insights can be gleaned from these studies. Firstly, Twitter indeed seems to be a valid source of data in terms of linguistic geography. Secondly, its use offers an advantage in that it allows a researcher to investigate a pool of data that is orders of magnitude greater than "normal" spoken language survey data. However, extracting useful data from a Twitter corpus is not without its issues, and will likely require the use of smoothing and other statistical tools. Nonetheless, studies using tweets as a data source for mapping linguistic features almost unanimously produce at least partially valid maps in terms of their correspondence to previously collected data on spoken language. Considering the multitude of studies that have been able to successfully use tweets as a source of linguistic data, the same should be true for the study of word-final *n*-deletion, a prime example of phonetic variation. As previously discussed, Van der Wal et al. (2012) report word-final *n*-deletion surfacing in written language centuries ago, so it stands to reason it could be appearing in modern-day tweets. Indeed, a study on spelling variation on Twitter by Van Halteren and Oostdijk (2012) reports cases of word-final *n*-deletion showing up in Dutch messages on Twitter. Therefore, examining tweets—especially their associated GPS metadata—appears to be a possibly promising approach to studying word-final *n*-deletion in the Netherlands and Flanders.

1.5 The present study

The goal of this study is to map the occurrence of word-final *n*-deletion in tweets from the Netherlands and Flanders while taking into consideration as many factors as possible that could influence its rate of occurrence. The main research question guiding this study is: what can the examination of Twitter data tell us about the degree to which the occurrence of word-final *n*-deletion is distributed across the Netherlands and Flanders, and to what degree do internal and external linguistic factors influence its prevalence? The utility of using Twitter data to map phonological features will also be investigated by attempting to answer the second research question: to what degree is the use of Twitter data useful in mapping individual phonological features, especially in terms of the quality of the results?

Considering previous work on word-final *n*-deletion, hypothesis 1 is that the results will indicate that word-final *n*-deletion will prove to be most prevalent in the North and North-East of the Netherlands and the westernmost third of Flanders. Even though there will probably be considerable differences between spoken language and tweets, it is likely that, as we are inspecting the same phenomenon but in a different context, the non-geographical factors that influence *n*-deletion in spoken language will influence it online as well, and in much the same way. Therefore, hypothesis 2 is that word type, right-hand context, gender, and time will influence *n*-deletion in the same way that they influence word-final *n*-deletion in spoken language, with the possible exception of right-hand context, considering Eisenstein (2015) reports phonological context seems to have a weaker influence on Twitter. Considering the literature on using tweets as data for mapping linguistic features, hypothesis 3 is that Twitter as a data source will prove fruitful in the case of word-final *n*-deletion in that the geographic data

generated from the data will likely match with existing knowledge of the distribution of word-final *n*-deletion to a considerable degree. It should be noted, however, that, contrary to this study, previous studies that employ Twitter data focus mostly on lexical features, which means this study will still, in some sense, be treading relatively new grounds. Therefore, whether hypothesis 3 will turn out to be congruous with reality remains less certain than is the case for hypothesis 1. The veracity of the hypotheses will be investigated by composing a large corpus of Dutch tweets, searching those tweets for relevant instances of word-final *n*-deletion, statistically analyzing those instances, and mapping the results. Specifically, our approach will be to map the degree to which users deviate from statistical predictions based on mostly non-user-related features. The features that will be included in the analysis and which are based on the previously discussed literature are visible in Table 1. After the analysis, the results will be compared to and evaluated based on existing knowledge of word-final *n*-deletion, and their implications for future research will be discussed.

Table 1: Discussed features and the status of their inclusion in this study

Internal factors		
Factor	*Studies*	*Inclusion*
Word type (part of speech and morphological complexity)	Ollevier (1959), Van Oss and Gussenhoven (1984), Van de Velde (1996), Van de Velde and Van Hout (1998), Van de Velde and Van Hout (2003)	Yes
Genre	Ollevier (1959), Koefoed (1979)	No
Style and/or formality	Koefoed (1979), Voortman (1994)	Partly (as standardness)
Right-hand context	Van Oss and Gussenhoven (1984), Van de Velde (1996), Van de Velde and Van Hout (1998); Van de Velde and Van Hout (2003)	Yes
External factors		
Gender	Koefoed (1979), Van de Velde and Van Hout (2001), Van de Velde and Van Hout (2003)	Yes
Age	Van Oss and Gussenhoven (1984), Van de Velde and Van Hout (2001)	No
Origin (geographical dimension)	Voortman (1994), Van de Velde (1996), Van de Velde and Van Hout (2001), Van de Velde and Van Hout (2003)	Yes
Time (diachronic dimension)	Van de Velde (1996)	Yes

2. Method

2.1 Corpus

The essential first step in any research project employing Twitter data is determining from where to source the data. Two options were apparent: make use of a pre-existing corpus or collect new data specifically for this study. While it is true that the majority of the authors of the previously discussed Twitter-based studies construct their own corpus, this approach would have come with several drawbacks. The most important of these is that Twitter's API only allows for the retrieval of 900 tweets every 15 minutes (Twitter Developer Platform, n.d.). This means that the collection of millions of tweets—a number that would be highly desirable, considering the advantages of using larger corpora—would have taken months. Secondly, the constraints imposed on the use of the Twitter data collection API are subject to the whim of its parent corporation. This meant that Twitter could have made the constraints on tweet collection more stringent during the data collection process, which would have unexpectedly interfered with the project. Therefore, the approach most suited to this study was making use of one or more datasets that were already available.

Fortunately, Van Halteren et al. (2018) reported using a large-scale corpus of tweets containing Dutch texts. The corpus in question is TwiNL by Tjong Kim Sang and Van den Bosch (2013). While originally released in 2013, as of this study, the current version of the corpus spans a period from 2010 up to and including 2019. In total, it contains 2,528,720,239 tweets authored by 64,955,636 users. Which such a wealth of data, the foremost concern was ensuring that all data used in the study was incorporated in an ethically defensible manner. In this case, that led to questions related to the public nature of Twitter. While, in theory, anyone is able to see public tweets, this does not mean that every tweet is authored with this fact in mind. Especially users with very few followers might consider their tweets more of a personal way of communicating than users with a large audience. It was essential, therefore, to devise a strategy to remove users from our dataset who had likely not intended for their tweets to be read and otherwise examined by a larger audience. Luckily, Twitter provided a built-in indicator of "publicness": hashtags. A user can prefix a word or other string of unbroken characters with a "#" to allow other users to easily retrieve tweets related to a certain topic, essentially publicly indexing a tweet. Filtering based on hashtags allows for the creation of a corpus that consists only of users that have deliberately chosen to expose their writing to the larger community. Removing all tweets without hashtags, however, would have inevitably distorted the results of the study, as it would have considerably narrowed its focus in terms of the genre of the studied tweets. When a user uses a hashtag, they not only allow for anyone to quickly find a specific tweet but also the user account associated with that tweet, and thus all other public tweets made by that user. Therefore, the issue was fixed by, instead of filtering out all tweets without hashtags, filtering out all users that had never used a hashtag in any of their tweets whatsoever, leaving only users who had at one point knowingly exposed their account to the public eye.

The as-of-yet unpublished TwiNT corpus (not to be confused with TwiNL), which was initiated by Stefan Grondelaers and collected by Jorrit Visser (2021), turned out to be a useful dataset here as well. This corpus consists only of tweets authored in and near the Netherlands and Flanders that contain one or more hashtags. However, this corpus is considerably smaller than TwiNL, containing 33,923,301 tweets by 1,043,804 users in total. In order to maximize the possible pool of data, combining TwiNT and TwiNL presented itself as the most expedient option. The approach for this study was to take the IDs for all users that appeared in TwiNT and to retrieve all other tweets by those users that appeared in the TwiNL corpus. This, combined with the fact that only derived data and not the tweets themselves were stored, left us with a large but ethically sound dataset of Dutch tweets. A request to construct and use the corpus as described here was submitted to Radboud University's Ethics Assessment Committee

Humanities (EACH) and was granted approval (see Appendix 2). This corpus formed the base of the study from thereon out.

2.2 Selecting users

To ensure that there was enough GPS information per user for accurate location determination, all users with fewer than 10 geotagged tweets were removed from the dataset. For each of the remaining users, all locations that lay beyond one standard deviation from that user's location's mean were removed. Subsequently, all users for whom less than two-thirds of the original GPS tags remained were filtered out of the dataset as well. For the remaining users, a "home location" was calculated by taking the mean latitude and longitude of their remaining locations. If a user's home location fell outside of the Netherlands or Flanders, they were also removed. While calculating a user's home location in this manner is likely to lead to an accurate result, Van Halteren et al. (2018) point out a possible caveat: users who only tweet or only enable their location while on vacation. Therefore, to eliminate these vacationers, all users who tweeted only during three months of the year or fewer were also removed from the dataset.

Another important factor in determining which accounts are relevant and which are not is the languages in which those users tweet. As this study is concerned with word-final *n*-deletion in Dutch, users who do not or rarely tweet in Dutch were also filtered out. There were multiple possible sources that could be used to determine the languages in which users tweeted. While the TwiNL corpus included its own attempt to determine the language of each tweet, this classification was not suitable: part of our data consisted of tweets from the TwiNT corpus, which does not provide such data. Both corpora did, however, provide the language Twitter itself had tried to determine for each tweet. Therefore, the decision was made to only take into consideration Twitter's own language classification. To ensure we had a dataset of sufficiently Dutch-writing users, only users who wrote in Dutch in two-thirds or more of their tweets were retained.

After filtering the dataset, 1,043,804 users with a combined total of 1,066,595,715 tweets remained in the dataset.

2.3 Detecting word-final n-deletion

After being filtered, the dataset was scanned for occurrences of word-final *n*-deletion and their non-deleted counterparts, firstly by employing regular expressions to find all words that could possibly constitute a case of either *n*-deletion or *n*-retention. After casting all text in lowercase and simplifying all emojis into one character (😀), the following regular expressions were used:

- For occurrences of word-final *n*-retention followed by a vowel:
  `\S+e+n+(?= +[aáàäeéèëiíìïoóòöuúùüyýÿ])`
- For occurrences of word-final *n*-deletion followed by a vowel:
  `\S+[aáàäeéèëuúùü]+h*(?= +[aáàäeéèëiíìïoóòöuúùüyýÿ])`
- For occurrences of word-final *n*-retention followed by a consonant:
  `\S+e+n+(?= +[qwrtpsdfghjklzxcvbnm])`
- For occurrences of word-final *n*-deletion followed by a consonant:
  `\S+[aáàäeéèëuúùü]+h* (?= +[qwrtpsdfghjklzxcvbnm])`
- For occurrences of word-final *n*-retention followed by a pause:
  `\S+e+n+(?= ?(?:$|\.|\,|\!|\;|\:|😀))`
- For occurrences of word-final *n*-deletion followed by a pause:
  `\S+[aáàäeéèëuúùü]+h*(?= ?(?:$|\.|\,|\!|\;|\:|😀))`

Before filtering non-relevant hits, relevant information was recorded for each hit:

- User ID: unique identification number associated with the author of the tweet in which the occurrence was found
- Tweet ID: unique identification number associated with the tweet in which the occurrence was found
- Occurrence string: the detected word in question (e.g., *dansen*).
- Standard version string: for occurrences of *n*-deletion, a non-*n*-deleted string was generated based on the occurrence string.
- Deletion: whether the string consists of a word with word-final *n*-deletion or not.
- Righthand context: whether the word in question is followed by a vowel, consonant, or pause (including emojis).
- Standard, non-standard, and emoji counts: absolute measures used to determine the (non-)standardness of a tweet. Each non-emoji word of the tweet in which the occurrence was found was checked against the OpenTaal wordlist (Stichting OpenTaal, 2020) in order to determine whether it was standard or non-standard. Emojis were counted separately from the standard and non-standard words.
- Part of speech: the standard version string of each occurrence was compared to the CELEX lexical database (Baayen et al., 1995) in order to determine all its possible part-of-speech categories. These categories were: (1) singular nouns, (2) plural nouns, (3) singular verbs, (4) plural verbs, and (5) spatial prepositions, adverbs, and adjectives. This information was stored as true/false for each category.
- Trema: whether the standard version of the occurrence ends on *ën* instead of *en*. This information is relevant when filtering out forms that have a vowel before *en* that could influence the status of that *en*. For example, a word like *alleen* (“alone”) ends on *een* (/en/) and should thus be filtered out, while a word like *reeën* (“roe deer”) still ends on (/ən/) and should therefore be maintained.
- Bigrams: bigrams in which the relevant word was found within the tweet.
- Trigram: the trigram in which the occurrence was found in the Tweet, with the occurrence in the middle position.
- Standard trigram count: number of times the trigram with the standard form of the occurrence occurred in all tweets made by users still in the corpus.
- Partial trigram count: number of times the trigram without specifying the middle part of the trigram occurred in all tweets made by users still in the corpus.
- Standard trigram frequency: the standard trigram count divided by the partial trigram count.
- Lengthening: whether the occurrence displayed lengthening of the word-final morpheme or not (e.g., *danseeeee* or *dansennnnn* instead of *danse*(*n*)).
- Date (day, month, year): on which date the tweet in which the occurrence was found was posted.
- Hour of the day: on which hour of the day the tweet in which the occurrence was found was posted.
- User’s tweet count: total number of tweets in our corpus posted by the user who posted the tweet in which the occurrence was found.
- User’s tweets per day: mean tweets per day of the user who posted the tweet in which the occurrence was found, computed over all dates between the user’s first tweet and their last.
- User’s possible gender: determination of each user’s gender was based on an internship project by Blonk (2021). She developed a classifier that categorized TwiNL users into one of four categories, namely male (certain), male (unsure), female (certain), and female (unsure). This classifier relied mostly on Dutch first name lists, with each user

being classified chiefly according to the full name they provided on their profile and their username. She reports that when manually inspecting a smaller sample for the purposes of error analysis, taking into account all four categories led to a recall of 0.925 and a precision of 0.800.

Using this data, false positives (i.e., detected occurrences that do not really constitute cases of *n*-deletion or their non-deleted equivalents) and occurrences that proved to be too ambiguous were removed. Firstly, all words ending on *-ie, -ien, -ui* were removed, as these were found not to be possible endings for *n*-deletion to occur. Then, using the OpenTaal wordlist (OpenTaal, 2022), cases of possible deletion which constituted words by themselves were removed, along with their standard version and other deleted variants. Any occurrence that occurred in dwyl's (2022) English word list was also removed to prevent any erroneous categorization of an Anglicism as a relevant occurrence, along with other deleted and non-deleted instances of that word in the dataset. Furthermore, words ending on *-een* that do not normally have a trema were also removed, as these words do not normally end on / ən/ in spoken Dutch. Lastly, the data was manually inspected by taking a random sample and by looking at the most frequent instances of (suspected) deletion in the data. Any word that did not constitute an actual case of deletion was removed from the data, along with what would have been its deleted and non-deleted variants, were it actually a case of deletion. This process of filtering left a total of 99,219,859 occurrences, 1,571,055 (1.58%) of which constituted *n*-deleted words. After filtering, the following additional variables were calculated for each occurrence:

- Proportion of deletion: the number of cases of deletion in the corpus out of all cases with the same standard form
- User deletion: how many times a user has deleted the *n* out of all detected *n*-deletions and *n*-retentions.
- Word frequency: how many times this word occurs in the set of *n*-deletions and *n*-retentions (includes both standard and non-standard versions of the word)

2.4 Pre-processing and statistical modeling

In order to assess the influence of the measured variables, the data was submitted to a number of statistical models. For the purposes of the analysis, a table was created in which each occurrence was represented as a separate row, with columns representing the features listed above in the list of calculated variables. Because was no major variation in deletion between days of the week and months of the year, these two variables were excluded from the set. The variables user ID, tweet ID, bigram string, trigram string, occurrence string, and standard version were excluded from the set, as these were not suitable for the purpose of statistical analysis. The variables dealing with right-hand context, grammatical category, user gender, year, and hour were converted into dummy variables with 0 or 1 for each possible value because, while they could technically be represented by one numerical variable, they do not constitute scalar variables. The scores dealing with the standard trigram frequency were replaced by their *z*-score because they exhibited a normal distribution. The variables dealing with standard trigram count, partial trigram count, user tweet count, and word frequency had their scores recalculated by taking the natural logarithm of that score plus 0.0000001 to avoid undefined values. This was done because their distributions were skewed toward higher values. Furthermore, for each column, its values were centered around 0 by subtracting their mean from all values, after which all values were divided by the column's highest *absolute* value, yielding columns with a minimum possible value of -1 and a maximum possible value of 1. This was done to increase the compatibility of the data with models that are sensitive to the scales of the features.

Following standard methodology for separating training and test data, we applied ten-fold cross-validation. In order to ensure occurrences originating from the same user were not divided across folds, the folds were generated based on the final digit of each occurrence's author's user ID, with IDs ending on 0 becoming fold 0, etc., yielding ten folds. Furthermore, to prevent information leakage, for each training set, the variables word frequency and the proportion of deletion were recalculated over the training set (i.e., the data excluding the test fold). If the frequency was zero for a certain occurrence, it was set to the mean frequency in the dataset over which the new frequencies were calculated.

Because cases of non-deletion were highly overrepresented when compared to deletions, the decision was made to train the model on downsized data. For each fold, the number of non-deletions was reduced to the number of deletions by randomly sampling non-deletions until the number of deletions and non-deletions was equal, after which the remaining non-deletions were discarded. This downsizing did not affect variables like the users' deletions, which were calculated over the non-downsized data.

The first of the employed statistical models was a logistic regression, which was conducted in the R programming language, using the *glm(family = "binomial")* function of its *stats* package for statistical modeling (R Core Team, 2020). This model was used because of its ability to predict binary categorical variables. In our case, that variable was whether or not a certain occurrence constitutes *n*-deletion or not. The use of a so-called Poisson regression, which can be used for discrete non-negative count data, was also considered. Using this type of regression does remain a viable alternative for further inquiries, though, on the condition the data is not downsized. This was determined to be a less optimal solution, however, as we were working with downsized data. A logistic regression model was fit to all features remaining in the dataset, with the exception of the user's deletion score, as this would have meant using the likelihood to delete to predict deletion, which would have constituted a contamination when calculating scores for mapping purposes. The same data was subjected to a random forest classifier as well, which also allows for the prediction of categorical variables. Specifically, use was made of the Python programming language (*The Python language reference*, 2019) and the package *scikit-learn* (Pedegrosa et al., 2011), with default settings for the classifier being retained. The use of a multilayer perceptron classifier was also considered, but an initial exploration would only yield overconfident error-prone predictions, after which it was decided to proceed with just the logistic regression and random forest classifier. After this, the same logistic regression and random forest classifier models were fit again, this time with the user's deletion score included, in order to assess whether the user's propensity to delete offers a significant contribution to the predictions. Finally, the two models were re-run with both the user's deletion score smoothed on the basis of their 2000 nearest neighbors (see the section below for further explanation). This was done in order to ascertain whether any possible geographic effect exists on the distribution of *n*-deletion in our data.

After running the analyses, for each model, the cut-off point yielding the best classification based on the raw prediction scores was calculated. This was done based on the false rejection rate (FRR) and false acceptance rate (FAR) by calculating the so-called equal error rate (EER), the point at which the FRR and the FAR are equal. The cut-off point was selected to approximate the EER to four decimals.

2.5 Mapping the data

After running the first statistical analysis, maps of the Netherlands and Flanders were created in order to visually represent any possible geographical influence on the distribution of *n*-deletion. This was done using R (R Core Team, 2020), specifically using the packages *SDMTools* (VanDerWal et al., 2014), *ggplot2* (Wickham, 2016), and *RANN* (Arya et al., 2019). Firstly, a deletion score for each user was computed per model, as described below:

1. For each raw value, which fraction of positive cases in our test data has a lower and which has a higher value is mapped, and which fraction of negative cases has a lower and which has a higher value is mapped as well. If for any of these fractions the result is 0, it is set to *1 / number of positive cases* if the fraction represents positive cases and to *1 / number of negative cases* if the fraction represents negative cases.
2. For each score, if the fraction of positive cases with a lower value is larger than the fraction of negative cases that have a higher value, then (with pos_{lower} denoting fraction of positive cases with a lower value and neg_{higher} the fraction of negative cases that have a higher value):
$$\log_2 \frac{pos_{lower}}{neg_{higher}}$$
Otherwise, the normalized score is equal to:
$$-\log_2 \frac{neg_{higher}}{pos_{lower}}$$
3. The ranges of the positive and negative values are linearly mapped separately by dividing by the most positive or most negative score respectively, so that the final normalized score is in the range from -1 to 1. By design, this yields scores in which the FAR is equal to the FRR, which means they are both equal to the EER.
4. Scores higher than 0 are categorized as deletions and scores lower than 0 are categorized as non-deletions.

It should be noted that these scores were computed based on the test data. While, in principle, they should have been done *leave-one-out*, with 99 million cases in each fold, any possible information leak should remain negligible at worst. After normalizing per fold, the same steps were carried out over all folds combined. As for the results of the random forest classifier, the predictions were heavily skewed toward 0 and 1. Values around 1 were not distinguishable because of the limited number of decimal digits. Fortunately, *scikit-learn* provides both deletion and non-deletion predictions, so at least one useful value per prediction is guaranteed. If the probability of deletion is smaller than 0.5, the adjusted score equals:

$$\log_2(prob_{del} + 0.000000001)$$

Otherwise, it is equal to:

$$-\log_2(prob_{nondel} + 0.000000001)$$

After this procedure, the adjusted random forest scores were fed into the same steps as described before as if they were raw logistic regression scores. After the scores for both the logistic regression and random forest classifier models were computed, scores per user ID were calculated for the purposes of mapping *n*-deletion. For every map, each user was represented by one point on the map, placed on their calculated home location. The mapped score was calculated using the following formula:

$$score = \ln \frac{deletions_{actual} + 1}{deletions_{predicted} + 1}$$

The reason additive smoothing with a value of 1 was used was to avoid any possible cases where there might occur a division by zero. A logarithmic scale was applied to increase the sensitivity to smaller differences in score at the lower end of the scale. After this, the scores were scaled to fit between 0 and 1:

$$score_{scaled} = \frac{score - score_{min}}{score_{max} - score_{min}}$$

After calculating the scores, the choice was made to represent each user as a separate point on a blank map of the Netherlands and Flanders. While some studies chose to aggregate scores by administrative division (e.g., Grieve et al. (2019)) or divide the area into equally sized rectangles (e.g., Donoso and Sánchez (2017)), the choice was made here to mirror the approach employed by Van Halteren et al. (2018). Mapping each user as an individual point allows for maps that accurately represent in which places users are most concentrated, minimizing any major possible visual effects of outliers in low-density areas. Because geographically close individual users sometimes differed greatly in scores, a smoothing method was applied, which was based on Van Halteren et al. (2018): for each user, a score was produced for *k* of its geographically closest neighbors, with maps being generated for *k* = 100 and 2000 in addition to an unsmoothed map. This score was calculated as follows: for each of the *k* nearest neighbors separately, the following subscore was calculated, with *distance* indicating the distance between that neighbor and the user for which the eventual score is being calculated, and $distance_{\max}$ indicating the distance between the user for which the eventual score is being calculated and the farthest selected neighbor:

$$subscore = \frac{-0.99 * distance}{distance_{max}}$$

After these subscores were calculated, the final score was computed by taking the mean of all subscores. For each map, a histogram with 10 bins was included to show the distribution of the scores. The bin which constituted the mode and its associated points on the map were colored black in order to establish a middle line, while scores in the bins below were colored blue and scores in the bins above were colored red, with a bin's color gaining in intensity proportionately to the bin's distance from the mode.

3. Results

3.1 Statistical models

The first logistic regression, which excluded the users' tendency to use deletions, yielded an EER of 0.1947. The second logistic regression, which included the users' tendency to use deletions, yielded an EER of 0.1175. The third logistic regression, which included the users' smoothed scores but no their tendency to use deletions, yielded an EER of 0.1993. The mean odds ratios by feature and model, calculated over the feature log odds and over all 10 folds, are visible in Table 3.1. Figures of the plots of the feature's odds per model instance are visible in Appendix 1. Odds for the features *hour_23, righthand_vowel,* and *year_2019* are shown as *NA* because they were excluded from the analysis due to that they would cause singularities—i.e., their contents are already fully described by other variables in the data.

*Table 3.1: Odds ratios per feature calculated over mean log odds over all 10 folds of the three logistic regressions by instance. Significance is indicated by *** for features that had a significance level of* p *= 0.0001 and ** for* p *= 0.001, while features without an asterisk always had a significance level* of p *> 0.05*

Feature	*Odds ratios 1*	*Odds ratios 2*	*Odds ratios 3*
(Intercept)	0.3296***	0.1452***	0.2853***
CELEX_A	2.2144***	2.1050***	2.2136***
CELEX_M	1.0873***	1.0925***	1.0873***
CELEX_N	1.1813***	1.0249	1.1822***
CELEX_V	1.8951***	2.2371***	1.8946***
CELEX_W	1.9590***	1.5880***	1.9590***
deleted_proportion	9.05×10^{10}***	7.72×10^{8}***	9.07×10^{10}***
hour_0	1.1138***	1.0502**	1.1137***
hour_1	1.2981***	1.1232***	1.2981***
hour_2	1.4437***	1.1647***	1.4439***
hour_3	1.5107***	1.2119***	1.5104***
hour_4	1.3492***	1.1047***	1.3488***
hour_5	1.0561	0.9759***	1.0572
hour_6	0.8398***	0.9118***	0.8395***
hour_7	0.8917***	0.9816	0.8914***
hour_8	0.6631***	0.7835***	0.6631***
hour_9	0.5906***	0.7165***	0.5907***
hour_10	0.6344***	0.7386***	0.6344***
hour_11	0.6676***	0.7636***	0.6677***
hour_12	0.7081***	0.7879***	0.7082***
hour_13	0.7069***	0.7998***	0.7068***
hour_14	0.7019***	0.7987***	0.7019***
hour_15	0.7215***	0.8152***	0.7216***
hour_16	0.7483***	0.8372***	0.7482***
hour_17	0.7885***	0.8746***	0.7882***
hour_18	0.8346***	0.9107***	0.8345***
hour_19	0.8122***	0.9118***	0.8122***
hour_20	0.8098***	0.9262***	0.8097***
hour_21	0.8573***	0.9623**	0.8574***
hour_22	0.9298***	0.9806	0.9294***

hour_23	NA	NA	NA
lengthening_yn	5.4189***	6.3556***	5.4195***
righthand_consonant	1.0035	1.0563***	1.0034
righthand_pause	2.5373***	2.8692***	2.5369***
righthand_vowel	NA	NA	NA
standard_trigram_count_log	0.0611***	0.0410***	0.0611***
standard_trigram_frequency_z	0.0919***	0.1095***	0.0919***
trema_yn	0.000	0.0000	0.0000
trigram_partial_count	0.5265***	0.6604***	0.5265***
tweet_emoji_count	1.64×10^{8}***	15.8943***	1.60×10^{8}***
tweet_nonstandard_count	6.38×10^{6}***	1.45×10^{8}***	6.41×10^{6}***
tweet_standard_count	0.0031***	0.0101***	0.0031***
user_gender_Mcertain	0.8174***	0.9375***	0.8154***
user_gender_Mmaybe	1.3075***	1.1597***	1.3017***
user_gender_Vcertain	1.1147***	1.2743***	1.1123***
user_gender_Vmaybe	1.5390***	1.4659***	1.5343***
user_tweetcount_log	0.0806***	0.7877***	0.0810***
user_tweetsperday	3.05×10^{10}***	2.4136**	1.24×10^{10}***
word_frequency_log	1.2570***	1.2072***	1.2568***
year_2010	8.9072***	6.2614***	8.9224***
year_2011	14.7696***	7.1664***	14.7743***
year_2012	15.7885***	6.3232***	15.8051***
year_2013	11.9104***	4.7325***	11.9200***
year_2014	5.4447***	2.9980***	5.4424***
year_2015	2.5940***	1.9927***	2.5914***
year_2016	1.7118***	1.4188***	1.7095***
year_2017	1.2820***	1.1240***	1.2800***
year_2018	0.9852	0.9005***	0.9838***
year_2019	NA	NA	NA
user_deletion	NA	4.80×10^{10}***	NA
user_smoothed	NA	NA	1.3292***

For the first logistic regression, almost all mean log odds had a coefficient of variation (i.e., the standard devation divided by the mean value) that was less than 0.1. The coefficients of variation for *user_gender_Vmaybe*, *hour_6*, *righthand_pause* and *year_2019* were higher than 0.1, but this was because the mean values for these variables were very close to 0. For *word_frequency_log*, the coefficient of variation was also higher (0.12), but here the fluctuation in value is likely due to its odds being the highest (and therefore more prone to fluctuations). For the second logistic regression, the only features with a coefficient of variation higher than 0.1 were *hour_0*, *hour_4*, *hour_5*, *hour_7*, *hour_22*, *user_tweetcount*, *year_2017*, and *year_2018*. For all of these features, except *user_tweetcount*, this was determined to be due to the fact that their mean values lie so close to 0. For *user_tweetcount*, the reason high standard deviation between folds remains unclear, although it is likely tied to the large size of its odds ratio. For the third logistic regression, the only features with log odds with a coefficient of variation higher than 0.1 were *righthand_consonant*, *user_gender_Vcertain*, *year_2018*, and *smoothed_score*. For all these scores, this high coefficient of variation was determined to be due to the fact that their mean values lie so close to 0.

The first random forest classifier model, which excluded the users' tendency to use deletions, yielded an EER of 0.1625. The second random forest classifier model, which included the users' tendency to use deletions, yielded an EER of 0.0844. The third random forest classifier model, which included the users' smoothed scores but no their tendency to use deletions, yielded an EER of 0.1625. The mean importances by feature, calculated over all 10 folds, are visible in Table 3.2.

Table 3.2: Feature importances of the random forest classifiers

Feature	*Importances 1*	*Importances 2*	*Importances 3*
CELEX_A	0.000911	0.000421	0.000659
CELEX_M	0.001592	0.000941	0.001041
CELEX_N	0.00019	7.81×10^{-5}	0.000261
CELEX_V	7.48×10^{-5}	6.56×10^{-5}	5.96×10^{-5}
CELEX_W	0.001094	0.000831	0.000733
deleted_proportion	0.117215	0.089307	0.113801
hour_0	4.47×10^{-5}	9.27×10^{-5}	9.38×10^{-5}
hour_1	0.000227	9.56×10^{-5}	0.000102
hour_2	3.58×10^{-5}	1.54×10^{-5}	7.11×10^{-5}
hour_3	1.06×10^{-5}	1.32×10^{-5}	3.46×10^{-5}
hour_4	4.73×10^{-6}	2.76×10^{-6}	5.34×10^{-6}
hour_5	3.27×10^{-6}	2.27×10^{-6}	2.32×10^{-6}
hour_6	2.70×10^{-6}	2.71×10^{-6}	2.12×10^{-6}
hour_7	7.93×10^{-6}	4.24×10^{-6}	7.50×10^{-6}
hour_8	1.01×10^{-5}	5.19×10^{-6}	9.76×10^{-6}
hour_9	6.87×10^{-5}	1.99×10^{-5}	4.52×10^{-5}
hour_10	1.07×10^{-5}	1.45×10^{-5}	3.41×10^{-5}
hour_11	6.24×10^{-6}	5.46×10^{-6}	7.14×10^{-6}
hour_12	4.17×10^{-6}	2.72×10^{-6}	3.43×10^{-6}
hour_13	4.25×10^{-6}	3.10×10^{-6}	2.98×10^{-6}
hour_14	5.53×10^{-6}	3.11×10^{-6}	3.20×10^{-6}
hour_15	4.62×10^{-6}	2.52×10^{-6}	3.81×10^{-6}
hour_16	2.92×10^{-6}	2.45×10^{-6}	3.78×10^{-6}
hour_17	3.71×10^{-6}	2.84×10^{-6}	2.75×10^{-6}
hour_18	4.06×10^{-6}	3.29×10^{-6}	3.12×10^{-6}
hour_19	2.97×10^{-6}	2.94×10^{-6}	3.97×10^{-6}
hour_20	7.55×10^{-6}	3.94×10^{-6}	5.04×10^{-6}
hour_21	5.51×10^{-6}	3.49×10^{-6}	3.11×10^{-6}
hour_22	1.01×10^{-5}	4.63×10^{-6}	8.67×10^{-6}
hour_23	2.08×10^{-5}	1.00×10^{-5}	4.51×10^{-5}
lengthening_yn	0.010888	0.006522	0.008731
righthand_consonant	0.002216	0.00187	0.004166
righthand_pause	0.011971	0.005864	0.014352
righthand_vowel	0.001301	0.000872	0.00167
standard_trigram_count_log	0.378005	0.251139	0.370387
standard_trigram_frequency_z	0.139049	0.111933	0.15023
trema_yn	3.67×10^{-5}	1.83×10^{-5}	3.10×10^{-5}
trigram_partial_count	0.150242	0.083449	0.154183

tweet_emoji_count	0.000493	0.000202	0.000482
tweet_nonstandard_count	0.014458	0.010851	0.013427
tweet_standard_count	0.062083	0.032518	0.056514
user_gender_Mcertain	0.001598	0.000882	0.001758
user_gender_Mmaybe	7.71×10^{-5}	2.26×10^{-5}	6.97×10^{-5}
user_gender_Vcertain	0.000349	0.000131	0.000378
user_gender_Vmaybe	0.000772	0.0003	0.000418
user_tweetcount_log	0.021916	0.011545	0.020528
user_tweetsperday	0.005694	0.004254	0.010166
word_frequency_log	0.007608	0.008907	0.011264
year_2010	1.31×10^{-5}	1.52×10^{-6}	8.43×10^{-7}
year_2011	0.005663	0.003347	0.005252
year_2012	0.01341	0.006468	0.011039
year_2013	0.005661	0.003558	0.006852
year_2014	0.001006	0.000283	0.000596
year_2015	0.006864	0.003881	0.006834
year_2016	0.008167	0.003545	0.008747
year_2017	0.010603	0.003648	0.009916
year_2018	0.00986	0.005208	0.009098
year_2019	0.008412	0.003159	0.005682
user_deletion	NA	0.3437	NA
user_smoothed	NA	NA	0.0001700

It should be noted that, although very low, when compared to the mean scores themselves, the standard deviations were very high. The mean coefficient of variation (i.e., the standard deviation divided by the mean value) for the first instance was 0.58, 0.62 for the second instance, and 0.57 for the third instance. In Appendix 2, all results for each feature are reported individually, including histograms and results from both logistic regressions and random forest classifiers.

3.2 Maps

This subsection contains the maps generated using the scores calculated on the basis of the model predictions and the users' deletion scores. A higher score indicates a higher degree of deletion. On the maps, the scores are divided into 10 bins with an equal range. Black indicates the bin with the most scores within it, while blue indicates scores lower than this bin and red scores higher than this bin. For each image, the top part displays a map of the Netherlands and Flanders with the users' locations indicated with points and their corresponding scores indicated through the described color scheme. The bottom section contains a histogram of the users' scores divided into ten bins, also colored according to the same scheme. Table 3.3 contains the maps generated based on data derived from actual model classifications by different smoothing levels. To assess the quality and validity of the maps resulting from the use of the smoothing algorithm, three additional maps were generated for the logistic regression predictions and a smoothing level of $k = 2000$ for which the order of the score column had been randomized. These maps are visible in Table 3.4.

Table 3.3: Maps of deletion score by model and smoothing level

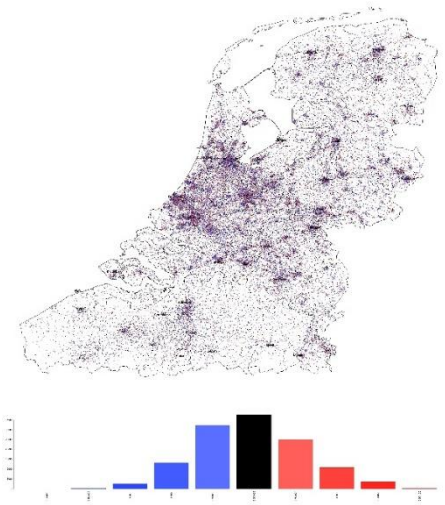
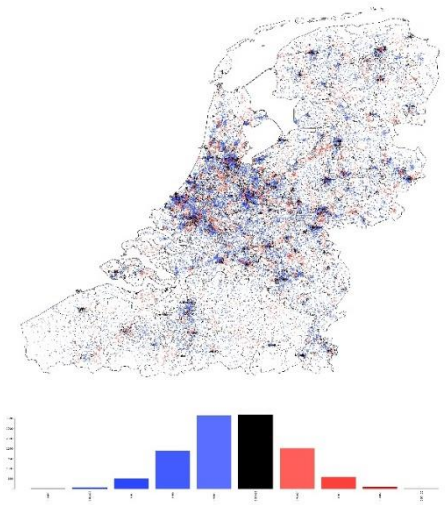
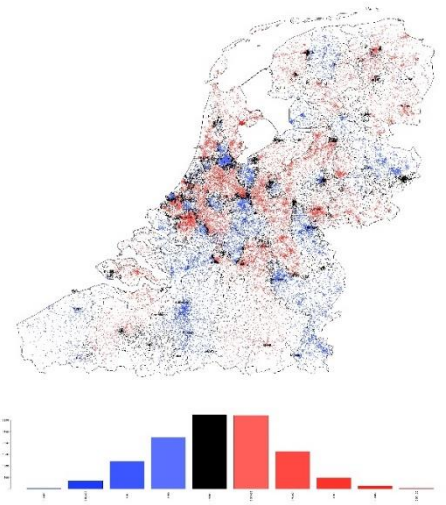
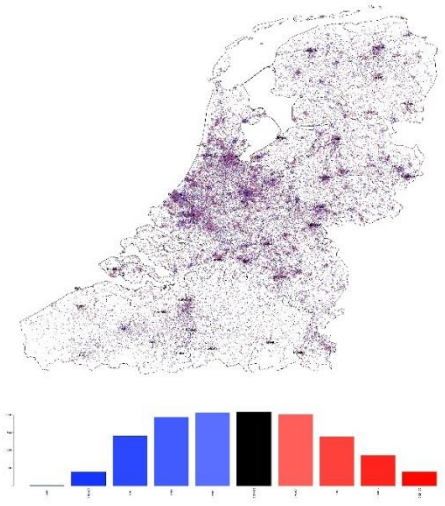
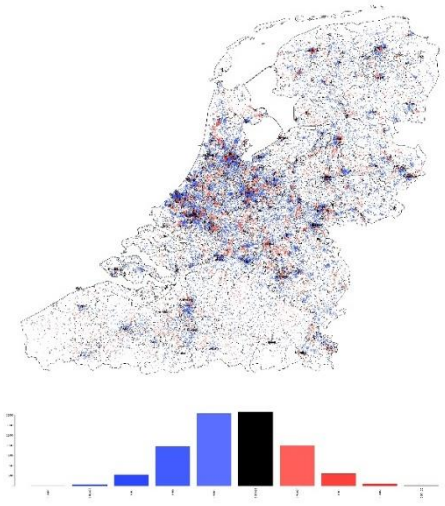
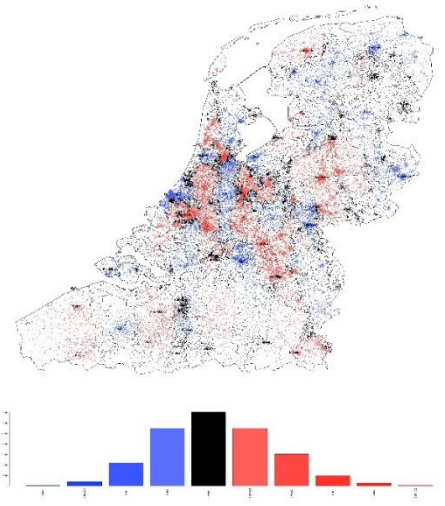
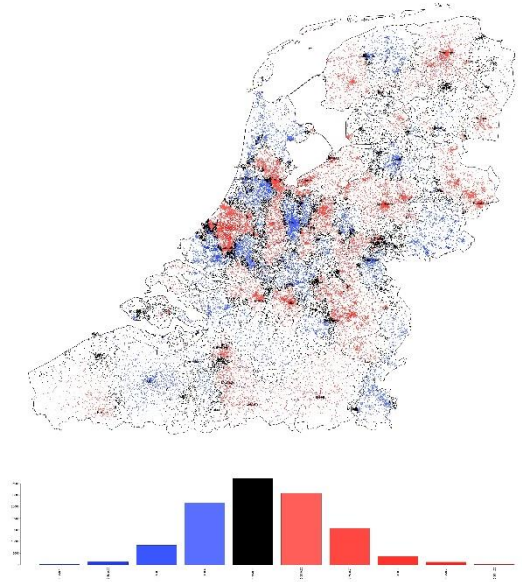
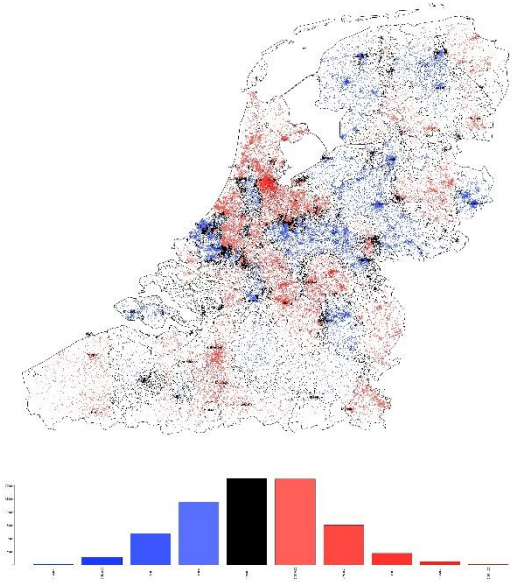
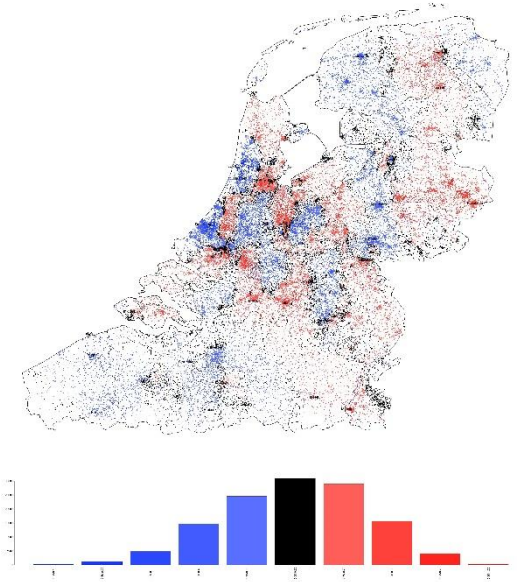
	Unsmoothed	*Smoothed (k = 100)*	*Smoothed (k = 2000)*
Logistic regression			
Random forest			

Table 3.4: Maps with randomized scores based on the logistic regression at k *= 2000*

Randomization 1	*Randomization 2*	*Randomization 3*
		

4. Discussion

The results have yielded a number of important insights, and it is imperative to discuss them thoroughly. Firstly, the generated maps will be discussed, after which the different model instances will be contrasted, individual features will be interpreted, the research questions will be answered, the limitations of this study will be discussed, and recommendations will be made for future research.

4.1 Maps

At first glance, the maps generated based on the scores fed into the smoothing algorithm seem to display areas that clearly differ in terms of how much *n*-deletion occurs in them. However, more closely inspecting the maps generated based on the predictions from the logistic regression and the random forest classifier, particularly those with $k = 2000$, reveals something remarkable: especially in the larger cities, like Amsterdam, the maps generated from the data provided by the two models contradict each other. One could be inclined to suspect that perhaps somewhere along the process, the polarity of the data might have been reversed. This is unlikely, however, as there are some regions between the maps that do match, such as the area around the city of Nijmegen. Instead, the contradictions between the maps seem to derive from the fact that the shapes and locations of the clusters that result from the smoothing are in many cases almost completely dissimilar in shape and size. Indeed, there appears to be almost no relation between what the maps are picturing.

This could have a number of different causes. On the one hand, it could be that the smoothing algorithm is simply pulling the regional patterns regions out of thin air, in that it is magnifying what essentially constitutes noise in the data into larger regions. On the other hand, it could be possible that there is some sort of major difference in the way the models predict *n*-deletion, in which case the issue does not lie with the smoothing algorithm. The most expedient way to assess the nature of this issue is to randomize the link between the observed scores and the user IDs and to map that data, as has been done for the logistic regression data above. If the algorithm truly is to blame, we would expect to see maps that mirror the non-random maps in terms of the pattern of clustering and coloration in general, while individual clusters would not line up with the non-random maps. If the issue lies with the model predictions themselves, then we would expect the maps of the randomized data to display completely different clustering patterns, if they even display clusters at all.

An inspection of the randomized smoothed maps tells us that they closely resemble the non-randomized maps in terms of the size and coloration of their clusters. This strongly indicates that the non-randomized maps are showing us nothing more than magnified noise. Therefore, it is highly likely that this issue is the root of the marked difference between the maps based on the logistic regression predictions and the maps based on the random forest classifier predictions, rather than some issue with one or both of the models. In order to confirm or nuance the likely inability of the smoothing algorithm to yield usable maps in our case, it is important we also compare the results of the statistical models trained on data that included the smoothed scores to the results of the other instances in which this data was not included.

4.2 Comparing model instances

The different logistic regression models on the one hand and the random forest models on the other will first be discussed separately. Comparing the first logistic regression model—i.e., the base model—with the second one—i.e., the model that includes the user's deletions as a feature—reveals a number of large differences. Firstly, the second model has a considerably more favorable EER at 0.118 when compared to the first model's EER of 0.195. Secondly, inspecting the odds ratios reveals that the features like the user's tweets per day and the number

of emojis used in the tweet have considerably lower odds ratios in the second model. While the implication of each feature's odds ratios will be discussed on a per-feature basis in the next section, this indicates that including the user's deletion behavior in the model has considerable predictive power. The feature *user_deletion* itself, when included, had a massively high odds ratio of 4.80×10^{10}. Considering this, the feature explains part of what the tweets per day and emoji count at first explained in the first model, and in a way that leads to a much lower EER. This result is not surprising, as this essentially only implies that including a user's deletion rate in the model allows for that user's deletions to be more accurately predicted. However, also having this second model allows us to contrast it with the third model, namely the model that includes the smoothed user scores. The logistic regression that included the smoothed user deletion score and not the unsmoothed one (like in the second model) was remarkably similar to the first model, both in EER and in the individual features' odds ratios. This points in the direction of the smoothed user scores (and therefore the resulting maps) not possessing any useful information on the deletion behavior of users. Otherwise, one would expect the model to improve considerably upon its inclusion.

Practically the same is true for the random forest models. The second random forest model which included user deletion scores performed considerably better than model one in terms of EER (0.1625 for model one and 0.0844 for model two), and the importance of deletion in the second model was very high (0.3437 out of 1). The third model, which included the smoothed user deletion scores, had the same EER as model 1 at 0.1625. Notably, the importances between model one and model three do differ more than in the case of the logistic regression models, but it should be noted that most have a relatively high coefficient of variation and lie close to 0.0. This means the importances are likely to vary between relatively similar models, so this is not a reason for concern.

What the comparisons of these models reveal is that the unsmoothed user deletion score *does* notably lower the EER, but that the smoothed user deletion score does not. This means that the smoothing has caused the scores to lose their explanatory power in terms of predicting *n*-deletion. Seeing as the smoothing was carried out on a geographical basis, this could have multiple explanations. It could be that either *n*-deletion is not spread geographically in written language on Twitter like it is in spoken language, or that our method was not sufficiently able to capture this geographical relationship. Considering that a method very similar to ours has proven to be successful in the past (Van Halteren et al., 2018), the former seems more likely. It should be noted that this certainly does not preclude the latter.

4.3 Factors

Discussing each feature entered into the models separately is likely to provide important insights. In this section, the odds ratios of the logistic regressions will be emphasized, and specifically those of the first model. The reason for not emphasizing the random forest importances more is that while they do say something about how much a feature contributed to the model, it does not show any directionality—in other words, it is not apparent from the importances whether an increase in the value of a feature leads the model to predict deletions with a higher or lower likelihood. Nonetheless, as the importances of each model add up to the value of 1, they still provide information on the role each feature had in the model in direct comparison with the other features, something which is not the case for the odds ratios. One issue is that certain features have very high importances others have very high odds ratios. This means that the scores for the features with smaller importances or odds ratios will be very small, meaning that the odds ratios and importances will be most useful during the interpretation of features with very high scores.

The first features to be subjected to a closer inspection are those that relate to the word type, i.e. the features starting with *CELEX*. Inspecting the histogram of *CELEX_A* (i.e., spatial

adjectives and prepositions) reveals that, when comparing the plot for all occurrences with the one containing only deletions, there is a slightly higher incidence of 1 for this feature when looking at the deletions only. The odds ratio in the first model is around 2.2, which also indicates that the presence of *CELEX_A* = 1 generally tends to lead to more deletion. The feature has a low importance score of about 0.001 in the first random forest model. *CELEX_N* (i.e., singular nouns) displays a slightly higher incidence of *CELEX_N* = 1 in the case of deletions only, although this difference does not stand out. The odds ratios are all around 1.2 for this feature, which points to this feature not being very predictive of deletion either positively or negatively. Its random forest importance score is also low (about 0.0002), which is lower than for *CELEX_A*. The histogram for *CELEX_M* (i.e., plural nouns) shows a much more pronounced difference between all cases and deletions only: the histogram for all cases shows the values of *CELEX_M* = 0 and *CELEX_M* = 1 distributed roughly 4:3, while the plot with only deletions shows that the features are much closer to being 1:1. This indicates a notably higher presence of *CELEX_M* = 1 relative to *CELEX_M* = 0 in the case of deletions only. However, this is contradicted by the feature's corresponding odds ratio, which hovers slightly under 1.1 for all three logistic regression models. Its importance score is about 0.0016, which does place it higher than *CELEX_A* and *CELEX_N*. The histogram for *CELEX_V* (i.e., singular verbs) does not display any difference between the plot containing all occurrences vs. the plot containing only deletions in terms of the relative distribution of the values of the feature, but, importantly, the odds ratio for *CELEX_V* are about 1.9, which would indicate a higher rate of deletion in the case of *CELEX_V* = 1. It does, however, have an importance score that is much lower than the other *CELEX* features, at around 7×10^{-5}. Lastly, inspecting the histogram for *CELEX_W* (i.e., plural verbs) shows that, in the case of deletions only, there is a slightly lower incidence of *CELEX_W* = 1, which would indicate that its presence generally leads to fewer deletions. This is contradicted by the odds ratio, which is roughly equal to 1.9 for models one and three, and 1.5 for model two. Its importance lies around 0.001, which is higher than *CELEX_A*, *CELEX_V*, and *CELEX_N*, but lower than *CELEX_M*.

When trying to interpret the results for word type, one runs into discrepancies between the histograms and the odds ratios for certain features, as has become apparent above. However, it is still possible to interpret the results in the places where these discrepancies do not manifest themselves. Generally, it seems that a word being a possible spatial preposition or adjective indicates a higher chance of deletion, which is also true for plural nouns and singular verbs. Singular nouns seem to be neutral in this respect, while plural verbs remain the most difficult to interpret. In our case, the category of "plural verb" is mostly congruous with the category of "polymorphemic verb." The literature on *n*-deletion in spoken language, however, makes the distinction between infinitives and participles, which we were not able to do. As shown by Van de Velde and Van Hout (2001; 2003), these two sub-categories can have significantly different rates of deletion, and it is possible that this plays a role in the ambiguity of the results for plural verbs. According to Van de Velde and Van Hout (2001; 2003), the order from least deletion to most deletion is monomorphemic verbs, monomorphemic nouns, polymorphemic finite verbs, spatial adjectives and prepositions, and polymorphemic infinitives. Therefore, there is not much agreement between deletions per word type in our data and deletions per word type in spoken Dutch.

The interpretation of *deleted_proportion*—the feature dealing with how much a word is deleted out of all cases—is relatively more straightforward. In its histogram, the main difference between all occurrences and deletions only is that the deletions-only plot has more scores in the higher 0.02-1.0 area of scores, while the plot with all occurrences had a relatively higher concentration in the far low end of scores (0.0-0.02). The deletions-only histogram does display a small peak near 1.0. This constitutes words that were deleted in 100% of cases that they (or their standard version) appeared. A manual inspection of the data reveals that these words are,

in large part, forms that are likely to be very rare (e.g., *Eucalpyten*, "Eucalyptus trees") and a small number of given names that should have been filtered out (e.g., "Thyrza") but which were missed during previous inspections. Generally, according to the histograms, the deletions exhibit a higher deletion proportion score. This is confirmed by the odds ratios in all three models, which range from about 8×10^{8} in the case of the second model (when the users' deletions scores are added) to about 9.05×10^{10} in the case of the first and third models. This is an enormously large value. While the directionality of the odds ratio does not surprise us, its sheer size does—we do not currently have an explanation as to why this and similarly large odds ratios in the results are so astronomically large. Its predictive power is also affirmed by the random forest importances, which all hover around 0.1 The results clearly exhibit that a higher *deleted_proportion* predicts a higher likelihood of deletion. Of course, this statement practically amounts to the truism "more deletion means more deletion," but its importance for this study does not lie in the implications of its inclusion and more in its usefulness in allowing for the creation of more accurate predictions and determining the relative value of the other features. This also ties into why *deleted_proportion*'s odds ratio decreases as *user_deletion* is added—some of the (likely user-specific) predictive power of *deleted_proportion* is more accurately described by *user_deletion.*

The next group of features is those that deal with the hour of the day. Rather than repeat the odds ratios and importances for each feature separately, a more general description of how their values change throughout the day will suffice. The odds ratio of *hour_0* starts off at about 1.1, moving to a peak of 1.5 at *hour_3,* after which it hits 0.83 at *hour_6* and hits the lowest point at about 0.59 at *hour_9.* It then slowly increases until it hits 0.92 at *hour_22,* with *hour_23* being selected as the reference category by the regression. The histogram does not have any directly visible discrepancies between all occurrences and deletions only. All hour features exhibited a low importance, ranging from about 3×10^{-6} to 0.0002 in the first model. If we focus on the odds ratios, we see that the chance of deletion is highest during the night, with a low point in the mornings, after which it slowly increases in the afternoon. A possible explanation for this is that Twitter accounts belonging to companies tweet mostly during the day, and those accounts are very unlikely to exhibit *n*-deletion in their tweets. It could also be that young people, who are more likely to stay up late, exhibit more *n*-deletion, causing the odds ratios to increase during the night. Notably, Hilte (2019) reports that young adolescents tend to use spelling variations more often than older adolescents expressing themselves online. Furthermore, if we look at the odds ratios for the second model, we notice that those values deviate much less from 1.0. This indicates that the influence of the time of day is tied to user-specific properties, which is likely to be at least in part the factor of age.

As for *lengthening_yn*, the most notable indicator of its relationship with deletions lies in the histogram. While lengthened words are exceedingly rare when taking into account all occurrences, they are much more clearly present when looking at deletions only. The odds ratio of the feature is also high at about 5.4. Its importance is also notable at about 0.01. These results clearly show that the presence of lengthening also increases the likelihood of that occurrence being deleted. There are multiple possible explanations. Firstly, a lengthened word is by definition non-standard orthographically. The same is true for *n*-deletion. In that sense, both types of spelling variation are connected. Secondly, the fact that *n*-deleted words end on a vowel might also play a role, as word-final vowels lend themselves more readily to lengthening in speech than consonants (compare *danseee* vs. *dansennn* for example). It is likely that both factors play a role in *lengthening_yn*'s positive predictive power in relation to *n*-deletion.

The features dealing with right-hand context do not display any difference between deletions and all occurrences. As for the odds ratios, right-hand vowel was the reference category, consonant has an odds ratio of about 1.0, and right-hand pause forms the aberration with an odds ratio of about 2.5. The consonant and vowel features have low importances, but

the pause feature has a notable importance of about 0.01. While the histogram shows no clear difference, the odds ratio for the pause feature would lead us to believe that the presence of pauses to the right of the word leads to more deletion. Here, too, there are multiple possible explanations. Firstly, according to Van de Velde and Van Hout (2003), the focus effect could play a role in how speakers realize word-final *n* before pauses. However, while this effect exists in spoken language, it is not likely to be at play here. In addition to the fact that Eisenstein (2015) reports that phonological factors play a smaller role on Twitter, we have already established that region likely does not influence the use of *n*-deletion on Twitter. Therefore, it seems unlikely that another aspect of spoken language, namely the importance of right-hand context, plays a part here. There is an alternative explanation that seems more likely. Because emojis following occurrences were also categorized as pauses, and the odds ratios for the emoji count feature (to be discussed in more detail later in this section), it could also be that the right-hand pause feature in essence simply captures the same type of relationship. This is further supported by the fact that there does not seem to be a major difference in how right-hand vowels and consonants are connected to deletion behavior—if factors relating to spoken language were at play here, one would expect them to behave similarly to how they behave in speech as well.

The next feature is the standard trigram count. While it is somewhat difficult to make out at first glance, the histogram does show some differences between all occurrences and only deletions. Most importantly, the values for the deletions only are more concentrated at the low end of the range of values in terms of frequency. The odds ratios, which approximate 0.06, also indicate that it predicts less deletion as it increases in value. Strikingly, the importance of the feature in the random forest models is very high, measuring at about 0.37 for the first model. This relationship is interpretable as showing that if a specific context (including the word itself) occurs more often, that means that word is less likely to be deleted. There is no straightforward explanation of this relationship, but there is a possibility. Seeing as a higher value of *word_count* indicates somewhat more deletion (discussed in detail below), the frequency of occurrence of the word itself cannot play a role here. Instead, the key must lie in the context. If we reverse how we characterized the relationship into the insight that rare contexts are more likely to be deleted than common ones, we reach a possibility: companies and bots are known to often tweet the same sentences many times. These accounts are less likely to exhibit non-standard language use, which is especially true for company accounts. Therefore, a possible explanation for that the fact that a higher standard trigram count means less deletion is that the results are skewed by company and bot accounts.

The feature dealing with the partial trigram account exhibits a similar relationship to *n*-deletion. Inspecting the histogram reveals that values on the lower end are more common when looking at just the deletions. The odds ratio is about 0.53, which is lower than for the standard trigram count, but this still indicates less deletion as the partial trigram count increases. The importance for this feature is high as well, at about 0.15. It would make sense to interpret this feature in a similar way to the standard trigram count, as we explained the relationship of that feature to *n*-deletion as being mostly dependent on its context and not on the word itself, which is exactly what the partial trigram count is supposed to capture. Here, too, then, the decrease of the likelihood of *n*-deletion occurring as the value of the feature increases could possibly be due to the influence of bot and company accounts.

The next feature to be examined is *standard_trigram_frequency*. The histogram reveals that, when taking into account deletions only, the occurrences seem to be more clustered around the low end of the range of the standard trigram frequency. This implies less *n*-deletion as the value of the feature increases. The odds ratios confirm this image, being around 0.09. The importance is considerable, just as with the other two trigram features, at 0.15. Here, too, interpreting this relationship is not a simple task. As we already know, for both the feature concerning standard trigram count and the feature concerning partial trigram count, the

likelihood of a deletion occurring decreases as the value of these features increases. The standard trigram *frequency* measures something different, however—namely, how likely a word is to occur in its specific trigram-based context. If a word almost always occurs in a certain trigram—which would lead to a high value for this feature and therefore likely less deletion—its context is probably considerably rare. As we know a higher word frequency leads to a higher rate of deletion, it is also to be expected that words that are rarer exhibit less deletion. In that sense, it is likely that *word_frequency* and *standard_trigram_frequency* are inversely related, which would explain the standard trigram frequency's relationship with *n*-deletion in our models.

The feature that is concerned with whether a word would occur with a trema in standard orthography is perhaps the most difficult to interpret. In its histogram, *trema_yn* = 1 is so rare it is not visible. Furthermore, the odds ratio for this feature in all three models is 0.000. This is in all likelihood an effect of its rarity, as this implies *trema_yn* = 1 never occurs with deletion. Therefore, the presence of a trema in standard orthography for a certain occurrence is so rare that this feature cannot be interpreted.

The next feature is the number of emojis in the tweet in which the occurrence was found. The histogram does not display any immediately noticeable difference between all occurrences and deletions only. However, on closer inspection, it seems that there seems to be a higher frequency of occurrences for deletions at the higher end of the range of the emoji count. It is important to note that we are dealing with the values of the feature after being subjected to a logarithm. Therefore, higher values are likely to have a larger effect than what the histogram might indicate. The odds ratios of this feature are also extremely large, measuring around 1.6×10^{8}. Its importance for the random forest models measures at a low 0.0005. In any case, however, a higher emoji count is certainly predictive of more *n*-deletions occurring. There are multiple possible explanations for this that do not necessarily preclude one another. Firstly, emojis have a distinctly expressive function. It is certainly possible for *n*-deletion in the case of Twitter to carry a similar meaning. Seeing as *n*-deletion is likely not firmly tied to *n*-deletion in spoken language, this different function could potentially play an important role. Furthermore, the emoji count could also be tied to the age of the speakers. As has already been established, an age effect for the use of *n*-deletion on Twitter is a real possibility, and it stands to reason that something similar is the case for the use of emojis, with younger users employing them more often. Notably, the odds ratio for this feature decreases to about 15, a very large reduction. This supports the effect of the feature being user-related. Furthermore, it could also be that the use of emojis is somehow correlated with non-standard language use in general. However, which of these possibilities is the most influential factor in the relationship between the emoji count and *n*-deletion is not apparent.

The next feature is the number of non-standard words counted in the tweet in which the deletion occurred. It is very similar to the previously discussed feature in that its histogram also shows a slightly higher relative frequency of deletions in higher values, while also being a logarithmic plot. Here, too, then, the histogram seems to point toward a positive correlation between non-standard word count and *n*-deletion. This is confirmed by a similarly high odds ratio of about 1.6×10^{6}. Its importance is relatively higher than the emoji count, however, at about 0.05. The relationship between the count of non-standard words and *n*-deletion seems rather obvious: as *n*-deletion in written language is itself an example of non-standard language use, it would make sense that it is correlated with other forms of non-standard language use. The only puzzling fact is that the feature has a higher odds ratio (around 1.5×10^{8}) for the second logistic regression. Why the inclusion of user deletion leads to a *higher* odds ratio for this feature remains inexplicable for now.

For the feature dealing with the number of standard words in the tweet in which an occurrence was found, the histogram exhibits a clear difference between all occurrences and

deletions only. Relative to all occurrences, the deletions taper off at a much lower value of *tweet_standard_count.* At approximately 0.003, the odds ratio reveals practically the same information. It is not surprising that this feature seems to be inversely related to the non-standard count in terms of their effect on *n*-deletion: just as more non-standard forms predict more *n*-deletion (which is itself a form of non-standard language use), it is logical that an increase in standard language use would predict a lower incidence of *n*-deletion.

The results for the features relating to the user's predicted gender are the next to be examined. As for the histogram, it is difficult to detect a difference between the plot of all occurrences and the plot of deletions only. The odds ratios, however, do reveal a contrast. Taken over the whole, the features predicting the user to be male led to lower odds ratios than the features predicting the user to be female. Notable is that the "maybe" categories both also scored higher than the "certain" categories. All features exhibited a relatively low importance score in the random forest models. An exact explanation for why the "maybe" categories led to more *n*-deletion is not certain, but it might be related to how Blonk (2021) categorized uncertain cases. In any case, these results imply that men are less likely to exhibit *n*-deletion on Twitter than women. While it is likely that the *n*-deletion patterns that have shown up in our Twitter data do not directly relate to *n*-deletion in spoken language, these results line up with what Van de Velde and Van Hout (2003) reported about *n*-deletion behavior of speakers from the Netherlands in particular. Seeing as users from the Netherlands are more heavily represented in our data than those from Flanders, this connection holds water. Hilte (2019) provides an interesting insight into the use of non-standard language online by adolescents in terms of a gendered difference: she reports that while male users tend to use more "traditional" non-standard language forms online, the female users use more expressive online-specific non-standard language. This lines up with the relationship between *n*-deletion and the incidence of emojis (the expressive function) and the idea that *n*-deletion online constitutes something separate from *n*-deletion in spoken language.

The next feature is the user's number of tweets. An investigation of the histogram indicates that, when only taking deletions into account, the plot is more heavily slanted toward the lower values—i.e., toward fewer tweets. The odds ratio, which hovers around 0.08, also indicates that a higher number of tweets for a user means less deletion. Interestingly, for the second regression, the odds ratio is much closer to 1, at 0.78. It is likely that the user's deletion score took on most of the explanatory power of this feature, as both are user-specific. The importance is also relatively notable at 0.02. The reasons for this relationship between tweet count and *n*-deletion could once again be related to company and bot accounts. These accounts, which are much less likely to *n*-delete, are also much more likely to tweet in general. While a possible effect of age could also be possible here, with Van Halteren (2021) reporting that many young Dutch people have left Twitter around 2014, and who therefore likely have a lower tweet count, the inclusion of the year in which the tweet was published should already account for this effect. Therefore, the influence of company and bot accounts remains the most likely factor here.

The other feature that deals with the number of a user's tweets, namely *user_tweetsperday*, displays the opposite relation to *n*-deletion. An inspection of the histogram reveals that for the plot with only deletions, there are relatively more values in the middle-to-high range. The odds ratio confirms this relationship at a staggering value of 3×10^{10}. Notably though, in the second logistic regression model, the odds ratio is only about 2. The importance scores were somewhat notable but not particularly high, ranging from 0.005 to 0.01 between the models. These results are interesting in that they are positively correlated with *n*-deletion, while the raw number of tweets was negatively correlated. The key to explaining this difference probably lies in how tweets per day were calculated—this feature takes into account only the window of time in which the user was active. Therefore, a user who tweeted intensely in during

a short window of time will probably have a high tweets per day score but a low total tweet count. That group could be youth users, who in large part left Twitter around 2015. However, as the year is also included as a factor, the model should have accounted for that effect. Looking at it from the opposite direction, however, is also useful: a user that tweets relatively often over a longer period of time does not delete as much. Again, company accounts could play a role here, as they are much less likely to use non-standard language and are highly active over long periods of time.

The next feature is the count of the standard version of the word in our corpus. Its histogram reveals no clear difference between all occurrences and deletions only. Some peaks occur in different locations on the *x*-axis, but there is no general pattern of difference between the two. The odds show that it is slightly predictive of more deletion at 1.25. Its importance ranges from 0.005 to 0.01. This means that words that are rarer are generally also *n*-deleted less, and that more common words are generally *n*-deleted more often. There are multiple possible reasons for this effect. Firstly, perhaps non-deleters, who tend to use more standard language, have also enjoyed a higher level of education, and, as a result, have a wider vocabulary. Those rare items are then also less frequently deleted because of their users. Secondly, it is possible that certain words are rare enough that they are never deleted at all. For those words, which have a very low frequency, the model would predict deletion to be very unlikely, causing the effect for word frequency. These two possibilities do not exclude one another, and there is no reason to believe only one of them is the cause of the effect.

We will now direct our attention to the features dealing with the year in which the tweet in which the occurrence was found was published. The histogram does not show any significant difference between all occurrences and deletions only. It does show the highest number of occurrences in general occurred between 2013 and 2016. The odds, however, are much higher for the earlier years (2010-2014), with a peak at 2011 and 2012 with an odds ratio of about 15. This decreases to about 5 for 2014 and 2 in 2015, finally going below 1 in 2018. 2019 was the reference category. The importances all range from about 0.005 to about 0.01, except for 2010. This is likely due to there being a low number of samples for that year in particular, as is visible in the histogram. If we direct our attention to the odds for the purposes of interpretation, we could also make a connection to young people leaving Twitter en masse around 2014, as this year and the decline in the likelihood of deletion line up. This is further supported by the fact that the total number of occurrences decreases after this year as well, pointing at a general decrease in activity on Twitter. As the number of tweets decreases, so does deletion. However, as there is no way to get the actual ages of the users in our dataset, the veracity of this explanation will remain unconfirmed for now.

The interpretation of the user's deletion score is relatively straightforward, as we have already discussed the second model specifically. Its histogram shows many more scores on the higher end in the case of deletions only, and its odds ratio is also astronomical at approximately 4.8×10^{10}. Its importance in the second random forest model is also very high at 0.34. The relationship here is clear, as a higher presence of deletions will naturally lead to a higher likelihood of deletions. One observation one can make, however, is that many deletions' users delete very often, as there is a higher frequency of values as the value of *user_deletion* increases when compared to the histogram with all occurrences. Inspection of the data shows that those users indeed delete very often, even in cases in which they have a very high total number of tweets. This indicates that word-final *n*-deletion is a persistent component of the language use of a large section of those who employ word-final *n*-deletion on Twitter. Considering deletions only, for the users closer to a score of 0.0 for *user_deletion*, in many cases their score approached 0.0 but never quite reached it, as the presence of one deletion would automatically indicate a score higher than 0.0. Those cases represent users that use word-final *n*-deletion very

sparingly. Therefore, while for one group of users word-final *n*-deletion seems to be a persistent part of their online writing, for another its use is limited only to certain instances.

Lastly, the user's smoothed score will be subjected to interpretation. The histogram does not show much of a difference between deletions only and all occurrences. The only difference is that there seem to be slightly more deletions at the higher end of the histogram, but not by much. If the smoothed score was truly predictive of the user's deletion behavior, one would expect to see more of a difference. The odds show that the smoothed score based on the regression predictions is slightly predictive of *n*-deletion at approximately 1.3, and the smoothed scores generated based on the random forest predictions exhibit a rather low importance score of about 0.0002. As discussed before when examining the third regression and random forest models, this effectively shows that the smoothed scores possess no major explanatory power anymore when compared to the raw user deletion scores. In other words, taking geography into account through the smoothing procedure, the link between the scores and *n*-deletion is effectively lost. This points to the smoothing algorithm essentially amplifying noise in the data to construct the regional groupings visible in the generated maps.

4.4 Answering the research questions and hypotheses

Now that all of the results have been examined, we have stable footing to answer the research questions and evaluate their concomitant hypotheses. The first research question was: what can the examination of Twitter data tell us about the degree to which the occurrence of word-final *n*-deletion is distributed across the Netherlands and Flanders, and to what degree do internal and external linguistic factors influence its prevalence? With regards to the first section of that question, namely the part dealing with using Twitter to investigate the geographical distribution of *n*-deletion, we hypothesized that results would indicate that *n*-deletion is most prevalent in the North and North-East of the Netherlands and the westernmost third of Flanders and that it is likely that online language use (like *n*-deletion) will mirror *n*-deletion in spoken language. Within the context of this study, this hypothesis must be discarded. The results show no connection between word-final *n*-deletion in spoken language and word-final *n*-deletion on Twitter in terms of geography. While it is possible that our method was not able to grasp that connection, previous studies like Van Halteren et al. (2018) have shown that methods very similar to ours have been able to capture other linguistic phenomena. Here, however, the generated maps seem to display nothing more than amplified noise, and the smoothed scores—which represent the geographical dimension of word-final *n*-deletion—do not notably improve model predictions, unlike the raw user deletion scores. Therefore, the examination of Twitter data was not able to tell us anything about the degree to which word-final *n*-deletion in spoken Dutch is distributed geographically. As for online *n*-deletion, the results tell us that it is likely not regionally bound.

The second hypothesis dealt with the remaining section of the first research question. It conjectured that other non-geographical factors that influence *n*-deletion in spoken language also influence it in the cases of our Twitter dataset. Firstly, according to our results, *n*-deletion on Twitter is least likely for singular nouns, followed by plural nouns, singular verbs, and spatial adjectives and prepositions. The position of plural verbs remains uncertain. According to the literature, monomorphemic verbs (equivalent to singular verbs in our study) are deleted least, then monomorphemic nouns (equivalent to singular nouns in our study), followed by polymorphemic finite verbs, spatial adjectives and prepositions, and finally polymorphemic infinitives (Van de Velde & Van Hout, 2003). While some similarities exist between the literature and our study, the orders differ from each other to a large degree. Secondly, our study reveals that *n*-deletion on Twitter is not influenced by whether they are followed by a vowel or a consonant, but it only occurs more if it is followed by a pause. It should be noted, however, that this was likely an effect of categorizing emojis as pauses. Therefore, it is likely that right-

hand context does not influence the occurrence of *n*-deletion on Twitter at all. On the contrary, in spoken language, *n*-deletion is certainly influenced by right-hand context. While differences exist between speakers, generally there is more deletion before consonants than vowels, and speakers are more likely to use their preferred pronunciation before pauses (Van Oss & Gussenhoven, 1984; Van de Velde, 1996; Van de Velde & Van Hout 1998; 2003). Thirdly, there seemed to be a higher degree of congruence between the influence of gender on *n*-deletion on Twitter and in spoken language. Generally, on Twitter, *n*-deletion seemed most likely for women and the least likely for men. According to Van de Velde and Van Hout, (2003), Dutch women delete more often than Dutch men, while the opposite is true for Flemish speakers. Considering the fact that the majority of our users have their location in the Netherlands, it would make sense for the Twitter data to match. However, there is also an alternative explanation—the exhibited gender pattern also matches with Hilte's (2019) study that found that male adolescents tend to use more traditional forms of non-standard language while female adolescents tend to use more expressive and online-specific forms. This expressivity also ties into *n*-deletion's frequent occurrence in tweets with emojis. Lastly, time was also studied as a factor. While older studies hypothesized that *n*-deletion was a change in progress (Van Oss & Gussenhoven, 1984), most relatively recent studies report this is not likely (Van de Velde, 1996). While our study found some change in the prevalence of *n*-deletion on Twitter, this is more likely linked to a change in social dynamics on Twitter than due to any ongoing language change in society. Considering all of these features, it seems that *n*-deletion in spoken language and on Twitter exhibit notable discrepancies, and it is likely that *n*-deletion on Twitter is not a projection of spoken language to online spaces but rather a separate phenomenon, similar to the expressive online language use as described by Hilte (2019).

The second research question was: to what degree is the use of Twitter data useful in mapping individual phonological features, especially in terms of the quality of the results? We hypothesized that Twitter as a data source would prove fruitful in the case of word-final *n*-deletion in that the maps generated on the basis of the data would likely match with existing knowledge of the distribution of word-final *n*-deletion to a considerable degree. In our case, this turned out not to be true. It is unlikely that the generated maps represent *n*-deletion as a phonological feature in spoken Dutch. Instead, it seems that the *n*-deletion that occurs in Dutch tweets constitutes a separate phenomenon with a distinctively expressive function. While our approach, therefore, did not prove useful in mapping phonological features, it did prove useful in studying online *n*-deletion through the features submitted to the models. While this study's original aim was not met, it has opened the door toward further study of this phenomenon.

4.5 Limitations

While it has yielded some intriguing results, this study was not without its limitations. Firstly, the selected phenomenon—namely, word-final *n*-deletion in Dutch—proved to be rather complex, with a large number of different factors influencing its occurrence. Future studies would do wisely in selecting a phenomenon whose influencing factors are already clearly established in the literature in case they are mostly focused on ascertaining the effectiveness of the use of Twitter data in studying spoken language phenomena. Otherwise, the perceived effectiveness of the data source might be influenced by the unclear nature of the factors that influence the use of the feature. Secondly, our data source needs to be discussed. The only limitation here is that we relied, in part, on an as-of-yet unpublished corpus, namely TwiNT. While it did prove to be a very useful data source, it is still inaccessible to other researchers. In the spirit of openness and reproducibility, we hope that this source will be released sometime in the near future. Thirdly, our set of mechanisms to detect cases of *n*-deletion or realization could have been more fine-tuned. An inspection of our data reveals that, even though it does not constitute a form of *n*-deletion but a non-standard lexeme, the word *fissa* ("party") was not

excluded from our set of deletions. This issue could have been avoided by a more thorough manual examination of the data. Another option would have been to select a limited number of *n*-deleted words and their standard forms and to search the corpus specifically for those, but this would have likely severely limited the size and scope of our dataset. Therefore, attempting to include as many different *n*-deleted and realized words remains the most desirable.

As far as the selection of the features themselves is concerned, it is true that some of the features indeed displayed a very high predictive ability. However, different choices could have still led to more accurate predictions. Firstly, the features that dealt with the occurrence of emojis, standard words, and non-standard words all represented the absolute number of incidences and not the proportion of words in the tweet that belonged to that category. A relative approach instead of an absolute one could have potentially further increased their already high predictive ability. Secondly, the non-standard count also included the *n*-deletions themselves. Obviously, this constitutes an information leak. However, this was missed during the data collection. Excluding the deletions and other occurrences from the (non-)standard counts would have been wiser in terms of preventing information leaks. Thirdly, the emoji count feature only included emoji characters and not so-called emoticons, like *:-D*. Seeing as they have very similar functions, adding emoticons to the emoji count would have made sense. On another note, more features could have also been added, most notably the length of the tweet. While there it is not likely that there is a connection between *n*-deletion and tweet length, there may very well be one anyway. Lastly, in addition to the continuous features, the categorical features were also scaled and centered. While this likely did not have an effect on this model, it does constitute an unnecessary transformation. As increasing the number of transformations of the data increases the risk of accidentally altering the data in some unexpected or otherwise undesirable way, this should have been avoided.

The modeling procedure was not without its issues either. Firstly, only a logistic regression and a random forest classifier were employed. Of course, there are many more types of prediction models that could have yielded more accurate predictions. While the scope of the study forced us to focus on the most promising model types, the inclusion of more models could have yielded more insightful results. Furthermore, only downsized data was used for the training phase of the models. While this was a useful procedure because of the rarity of *n*-deletion in our dataset, downsizing is not the only option to tackle this issue, as other types of models are more adept at handling this issue. The best way to handle imbalanced data, however, depends greatly on the specifics of the study (for a survey and discussion on this issue, see Kaur et al., 2020). In addition, the feature importances of the random forest results, while somewhat useful in determining which features had a very large influence on the predictions, were less useful for interpreting the features with a relatively small impact. This was also adversely affected by the high degree of variances in these features between folds. This, again, means other models could also have been considered that would have avoided the issue of very small importances. Something similar was the case with the odds ratios for the most predictive features in the logistic regression models, which in some cases approached dozens of billions in size. Because these odds ratios were so unexpectedly massive, there is no saying how these features have affected the features with lower odds ratios. Running instances of the models with the outliers in terms of importance and odds ratios removed could have provided a clearer image of the features with lower importance and odds scores. Lastly, one possible issue is that the models were provided with a large number of different features. While it is desirable to take as many factors into account as possible, it is likely that entering fewer features would have made the interpretation of those remaining features more straightforward and less error-prone. Notwithstanding these issues, the models that we did choose to employ still yielded very useful information.

A different approach to the mapping of the results could also have been taken. Firstly, we used a *k*-nearest-neighbors smoothing approach. This brings with it the possibility that more densely populated areas pull in less densely populated areas into their clusters, perhaps disproportionately so. The only way to figure this out, however, is to incorporate other clustering methods to see if there are any noticeable differences between the approaches. Secondly, in this study, each user was mapped as a separate point. This was useful for showing the density of the users, but we could have also employed a strategy like the aggregation of users by either equally sized rectangles on the map or by administrative division (like municipalities). However, considering that previous studies like Van Halteren et al. (2018) have employed a similar approach successfully, whether those other approaches would have proven useful remains to be seen.

All of the suggestions made above are recommendations for future studies to take into account. We also intend to explore word-final *n*-deletion on Twitter and our data further and will focus on tackling this study's limitations. This includes attempting to use different types of models, ameliorating feature-specific issues, and being more selective with those features. Beyond that, we also recommend that future studies delve into other phonological features, possibly in different languages as well, to further evaluate the use of Twitter-sourced data in sociolinguistic research. This study also showed a promising influence of gender specifically on online language use, and sociolinguistic studies beyond those that focus on mapping particular features would do well to attempt to take this factor into account.

5. Conclusion

The goal of this study was to map the occurrence of word-final *n*-deletion in tweets from the Netherlands and Flanders while taking into consideration as many factors as possible that could influence its rate of occurrence. The main research guiding the study was: what can the examination of Twitter data tell us about the degree to which the occurrence of word-final *n*-deletion is distributed across the Netherlands and Flanders, and to what degree do internal and external linguistic factors influence its prevalence? We hypothesized that *n*-deletion on Twitter would mirror *n*-deletion in spoken Dutch in that it would be most prevalent in the North and North-East of the Netherlands and the westernmost third of Flanders. This hypothesis could not be confirmed. While this could possibly be due to the methods employed, it is likely that *n*-deletion in spoken Dutch and *n*-deletion on Twitter constitute separate phenomena, with the latter not being regionally bound and having a distinctive expressive function. We also hypothesized that non-geographical factors that influence *n*-deletion in spoken language would also influence *n*-deletion on Twitter in the same way. However, as with the answer to the first hypothesis, the results indicate that *n*-deletion in spoken language and on Twitter are mostly influenced differently by the same factors. The secondary research question was: to what degree is the use of Twitter data useful in mapping individual phonological features, especially in terms of the quality of the results? Here, our hypothesis was that the use of Twitter data would prove fruitful in the mapping of phonological features (word-final *n*-deletion in our case). While the results showed that specifically mapping these features was not able to get at the distribution of *n*-deletion in spoken Dutch, as this likely constitutes a separate phenomenon. At the same time, the results did prove to be useful for the study of online *n*-deletion, which includes the fact that it is likely not regionally bound. However, they were also unusual in that the odds for some of the features included were astronomically high, something which could not be explained. Future studies delving into word-final *n*-deletion and similar ones should devote special attention to feature selection and the modeling and mapping procedures. Nonetheless, this study constitutes the first report of *n*-deletion as an expressive online phenomenon and lays the foundation for future studies on this topic.

References

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In *WWW'18: Proceedings of the 2018 World Wide Web conference* (pp. 1125–1134). International World Wide Web Conference Committee. https://doi.org/10.1145/3178876.3186011

Arya, S., Mount, D., Kemp, S. E., & Jefferis, G. (2019). *RANN: Fast nearest neighbor search using L2 metric* (Version 2.6.1) [R package].

Baayen, R. H., Piepenbrock, R., & L. Gulikers. (1995). *The CELEX lexical database* (Dutch version 3.1) [CD-ROM]. Linguistic Data Consortium.

Blonk, A. (2021). *Trek taal van Twitter om taal van Twitter te tracken* [Unpublished internship report].

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words! Linguistic style accommodation in social media. In *WWW '11: Proceedings of the 20th international conference on world wide web* (pp. 745–754). Association for Computing Machinery. https://doi.org/10.1145/1963405.1963509

Donoso, G., & Sánchez, D. (2017). Dialectometric analysis of language variation in Twitter. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)* (pp. 16–25). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1202

Doyle, G. (2014). Mapping dialectal variation by querying social media. In *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics* (pp. 98–106). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1011

dwyl. (2022). *List of English words* [Word list]. https://github.com/dwyl/english-words

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics, 19*(2), 161–188. https://doi.org/10.1111/josl.12119

Eisenstein, J. (2017). Identifying regional dialects in on-line social media. In C. Boberg, J. Nerbonne, & D. Watt, *The handbook of dialectology* (pp. 368–383). Wiley Blackwell. https://doi.org/10.1002/9781118827628.ch21

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS ONE*, 9(11), 1–13. https://doi.org/10.1371/journal.pone.0113114

Goeman, T. (2001). Morfologische condities op n-behoud en n-deletie in dialecten van Nederland. *Taal en Tongval themanummer, 14*, 52–88.

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PLOS ONE, 9*(1), 1–6. https://doi.org/10.1371/journal.pone.0112074

Gray, T. J., Danforth, C. M., & Dodds, P. S. (2020). Hahahahaha, duuuuude, yeeessss! A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings. *PLOS ONE, 15*(5), 1–27. https://doi.org/10.1371/journal.pone.0232938

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence, 2*(11), 1–18. https://doi.org/10.3389/frai.2019.00011

Grieve, J., Nini, A., & Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics, 46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Hilte, L. (2019). *The social in social media writing: The impact of age, gender and social class indicators on adolescents' informal online writing practices* [Doctoral dissertation, University of Antwerp].

Ilbury, C. (2019). “Sassy Queens”: Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics, 24*(2), 245–264. https://doi.org/10.1111/josl.12366

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using “Black Twitter.” *American Speech, 90*(4), 403–440. https://doi.org/10.1215/00031283-3442117

Kaur, H., Pannu, H. S., & Malhi, A. K. (2020). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys, 52*(4), 1–36. https://doi.org/10.1145/3343440

Koefoed, G. A. T. (1979). Paradigmatische invloeden op fonetische processen. In T. Hoekstra & H. van der Hulst (Eds.), *Glot special: Morfologie in Nederland* (pp. 51–70). Vakgroep Nederlands, Rijksuniversiteit Leiden.

Leupenius, P. (1958). *Aanmerkingen op de Neederduitsche taale en naaberecht.* Wolters. (Original work published 1653)

Ljubešić, N., Miličević Petrović, M., & Samardžić, T. (2018). Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography, 6*(2), 100–124. https://doi.org/10.1017/jlg.2018.9

Lüdecke, D. (2022). *sjPlot: Data visualisation for statistics in social science* (Version 2.8.12) [R package].

Ollevier, P. (1959). *Apocope van -n na toonloze vocaal bij de niet-adnominale woorden in Zuid-Nederland.* Unpublished manuscript.

Pauwels, J. L. (1969). Eind-n na toonloze vocaal in Zuid-Nederland. *Taal en Tongval, 21,* 216–218.

Pedegrosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.3) [Software]. R Foundation for Statistical Computing.

Shoemark, P., Kirby, J., & Goldwater, S. (2017). Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation* (pp. 59–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4908

Stichting OpenTaal. (2020). *Nederlandse woordenlijst* (Version 2.20) [Word list]. https://github.com/OpenTaal/opentaal-wordlist

Tatman, R. (2015). #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle of the University of Victoria, 25*(2), 97–108.

The Python language reference: Version 3.8. (2019). Python Software Foundation.

Tjong Kim Sang, E., & Van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal, 3*, 121–134.

Van de Velde, H. (1996). *Variatie en verandering het gesproken Standaard-Nederlands (1935–1993)* [Doctoral dissertation, Catholic University Nijmegen].

Van de Velde, H., & Van Hout, R. (1998). Dangerous aggregations: A case study of Dutch (n) deletion. In C. Paradis et al. (Eds.), *Papers in Sociolinguistics* (pp. 137–147). Éditions Nota Bene.

Van de Velde, H., & Van Hout, R. (2001). N-deletion in reading style. In H. de Hoop & T. van der Wouden (Eds.), *Linguistics in the Netherlands 2000.* John Benjamins Publishing Company. https://doi.org/10.1075/avt.17.20van

Van de Velde, H., & Van Hout, R. (2003). De deletie van de slot-n. *Nederlandse Taalkunde, 8*(2), 93–114.

Van der Wal, M., Rutten, G., & Simons, T. (2012). Letters as loot: Confiscated letters filling major gaps in the history of Dutch. In M. Dossena & G. Del Lungo Camiciotti (Eds.), *Letter writing in Late Modern Europe* (pp. 139–161). John Benjamins Publishing Company.

Van Halteren, H., & Oostdijk, N. (2012). Towards identifying normal forms for various word form spellings on Twitter. *Computational Linguistics in the Netherlands Journal, 2,* 2–22.

Van Halteren, H., Van Hout, R., & Roumans, R. (2018). Tweet geography: Tweet based mapping of dialect features in Dutch Limburg. *Computational Linguistics in the Netherlands Journal, 8,* 38–162.

Van Halteren, H. (2021). *Pitfalls in tweet-based variation studies* [Conference presentation]. Twitter as a laboratory for language variation and change: New opportunities for social media-based sociolinguistic research (NWAV49), Austin, TX.

Van Heule, C. (1953). *De Nederduytsche grammatica ofte spraec-konst.* Wolters. (Original work published 1625)

Van Oss, F., & Gussenhoven, C. (1984). De Nederlandse slot-n in het nieuws. *Gramma, 8*(1), 37–45.

VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L., & Storlie, C. (2014). *SDMTools: Species distribution modelling tools* (Version 1.1–221) [R package].

Visser, J. (2021). *Opbouw van een Nederlands Twittercorpus met Twint* [Unpublished internship report].

Voortman, B. (1994). *Regionale variatie in het taalgebruik van notabelen: Een sociolinguïstisch onderzoek in Middelburg, Roermond en Zutphen* [Doctoral dissertation, University of Amsterdam]. IFOTT.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [R package].

Willis, D. (2020). Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa, 5*(1), 1–33. https://doi.org/10.5334/gjgl.1073

Appendix 1: Plots of feature odds

This appendix contains odds ratio plots for all three logistic regression models, created using the sjPlot package (Lüdecke, 2022) in the R programming language (R Core Team, 2020).

Figure 1: Odds plot for the first model instance with all features included. The features user_tweetsperday and deleted_proportion *had odds that were too large to be displayed.*

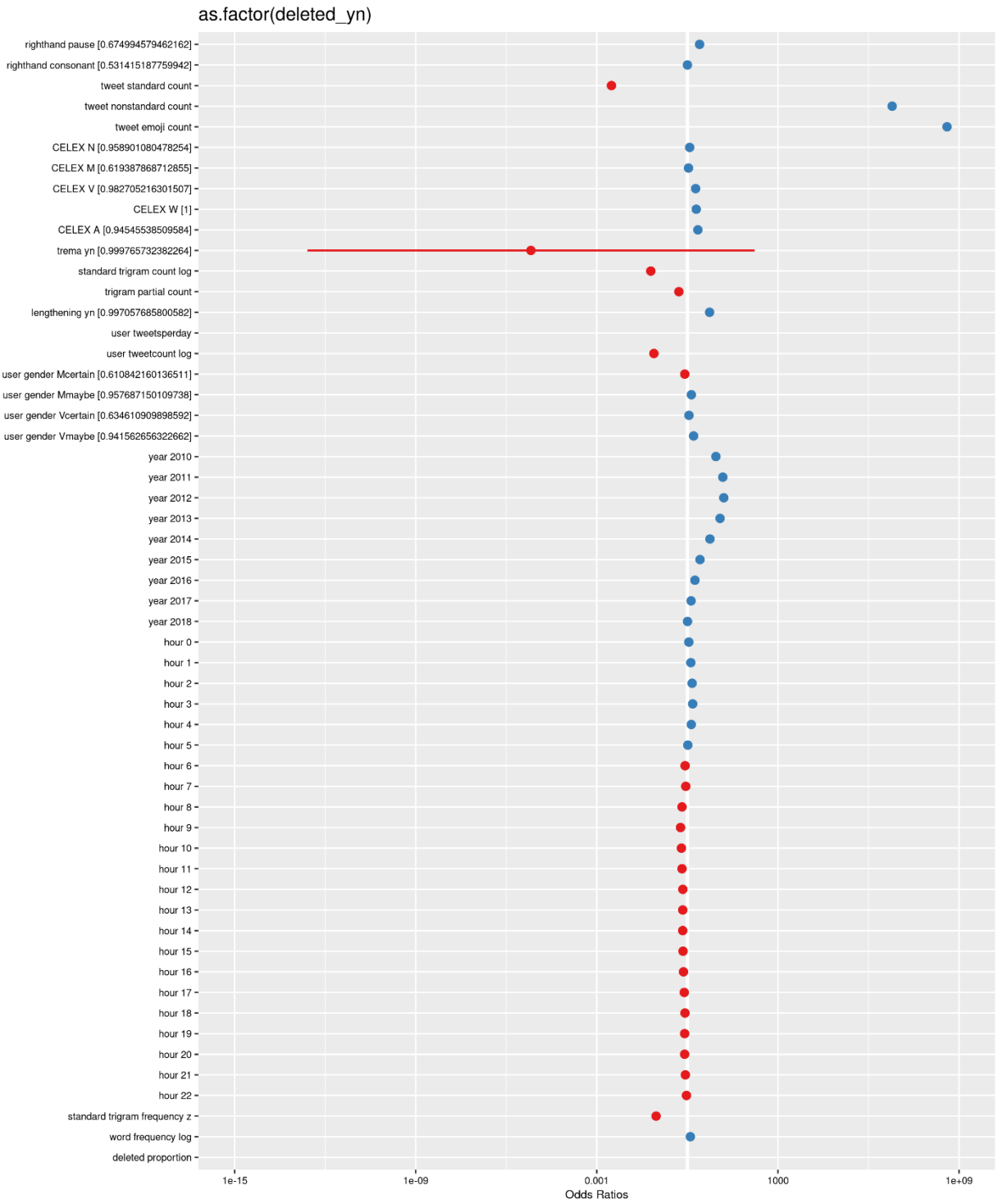


Figure 2: Odds plot for the first model instance with the features with the highest odds removed.

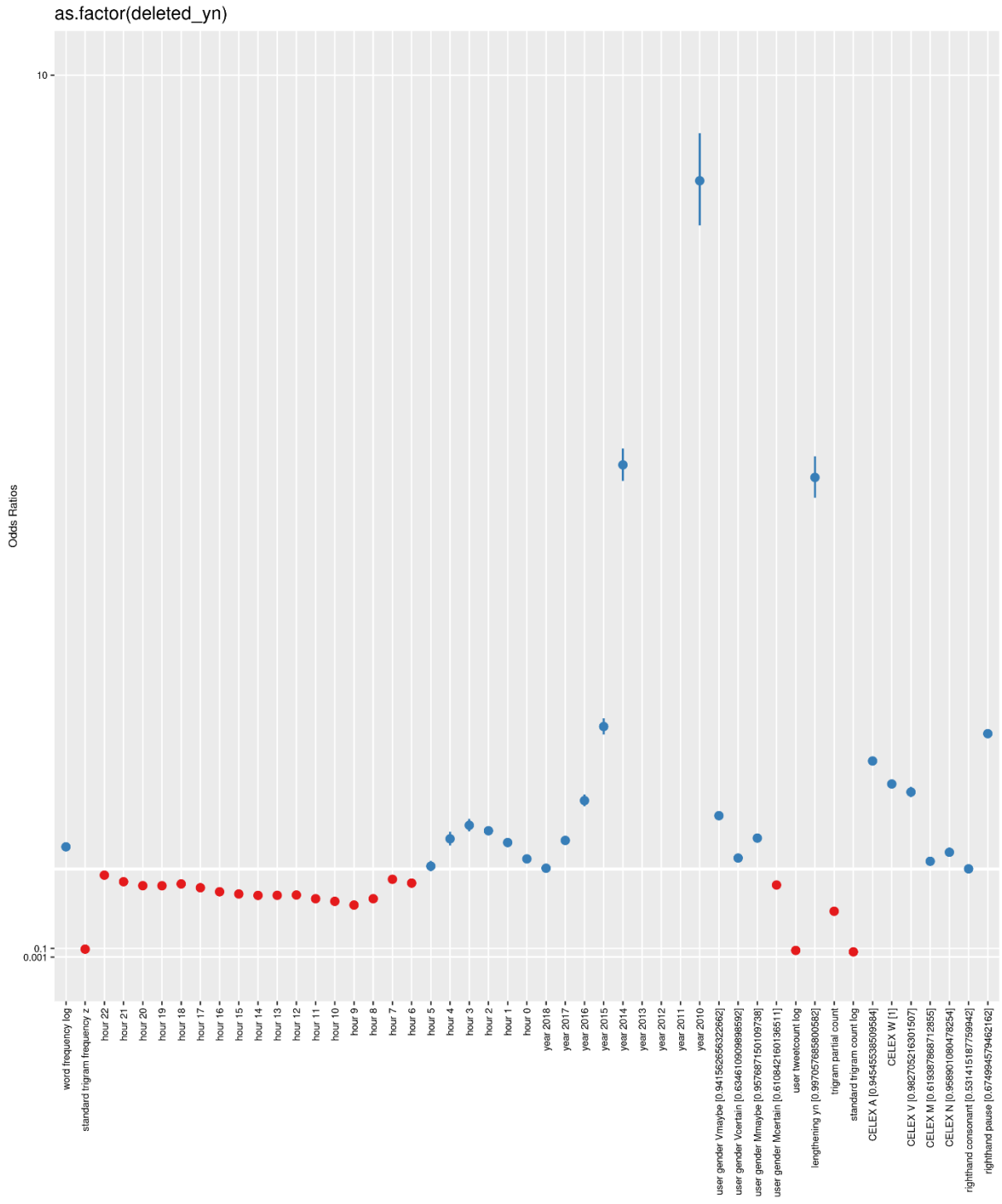


Figure 3: Odds plot for the second model instance with all features included. The features tweet_nonstandard_count, user_deletion *and* deletion_proportion *had odds that were too large to be displayed.*

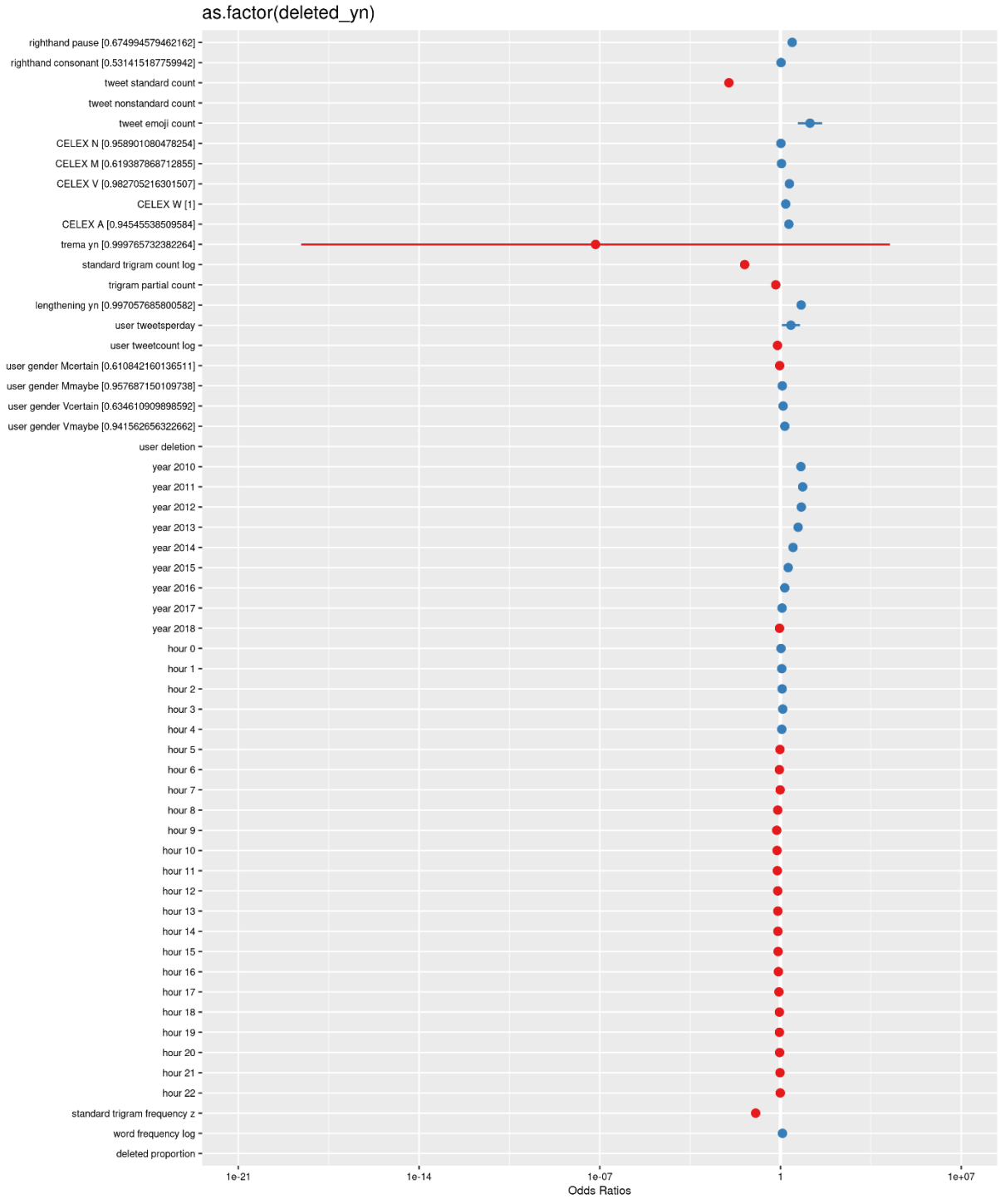


Figure 4: Odds plot for the second model instance with the features with the highest odds removed.

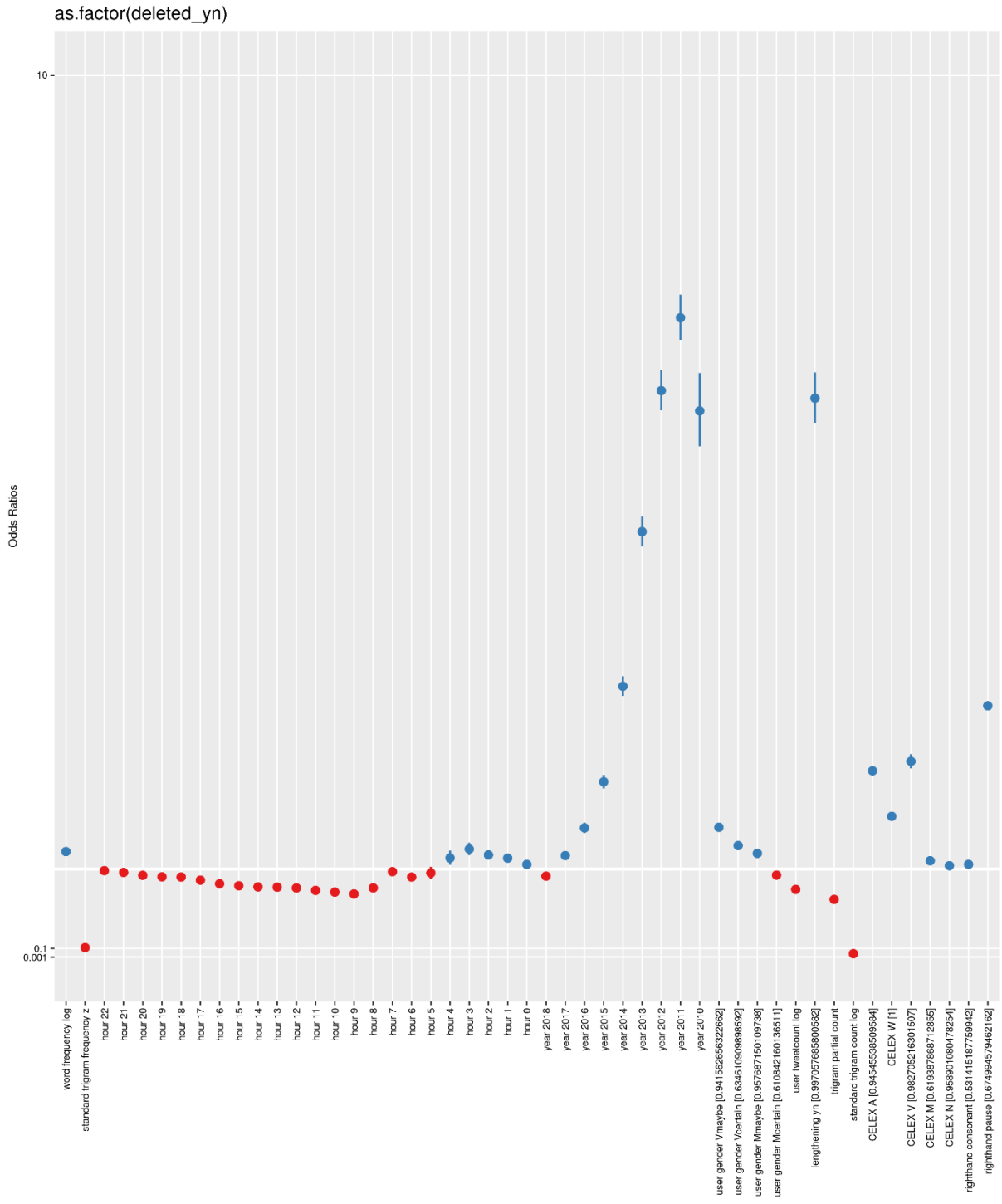


Figure 5: Odds plot for the third model instance with all features included. The features tweet_nonstandard_count *and* deletion_proportion *had odds that were too large to be displayed.*

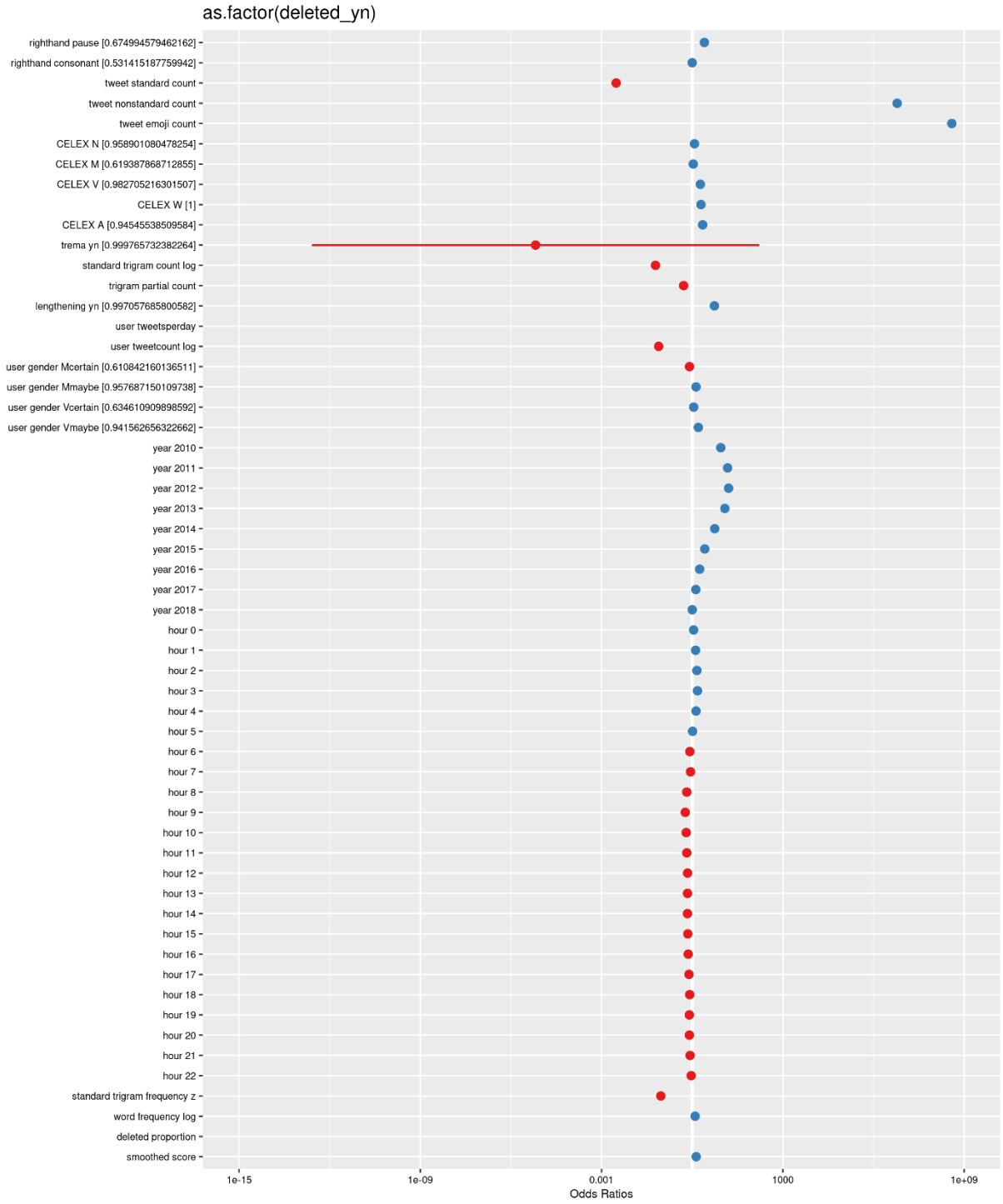
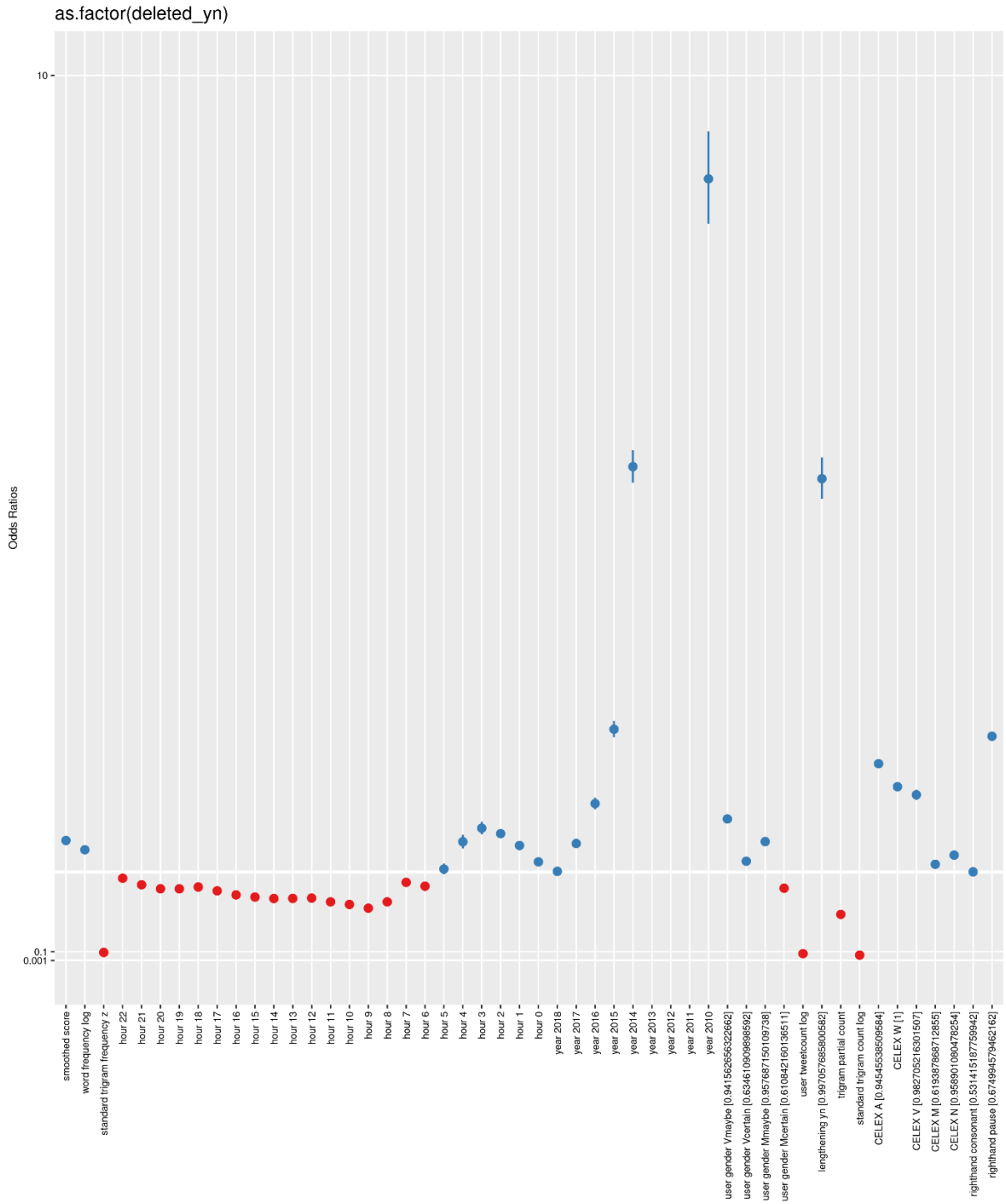


Figure 6: Odds plot for the third model instance with the features with the highest odds removed.



Appendix 2: Results per model feature

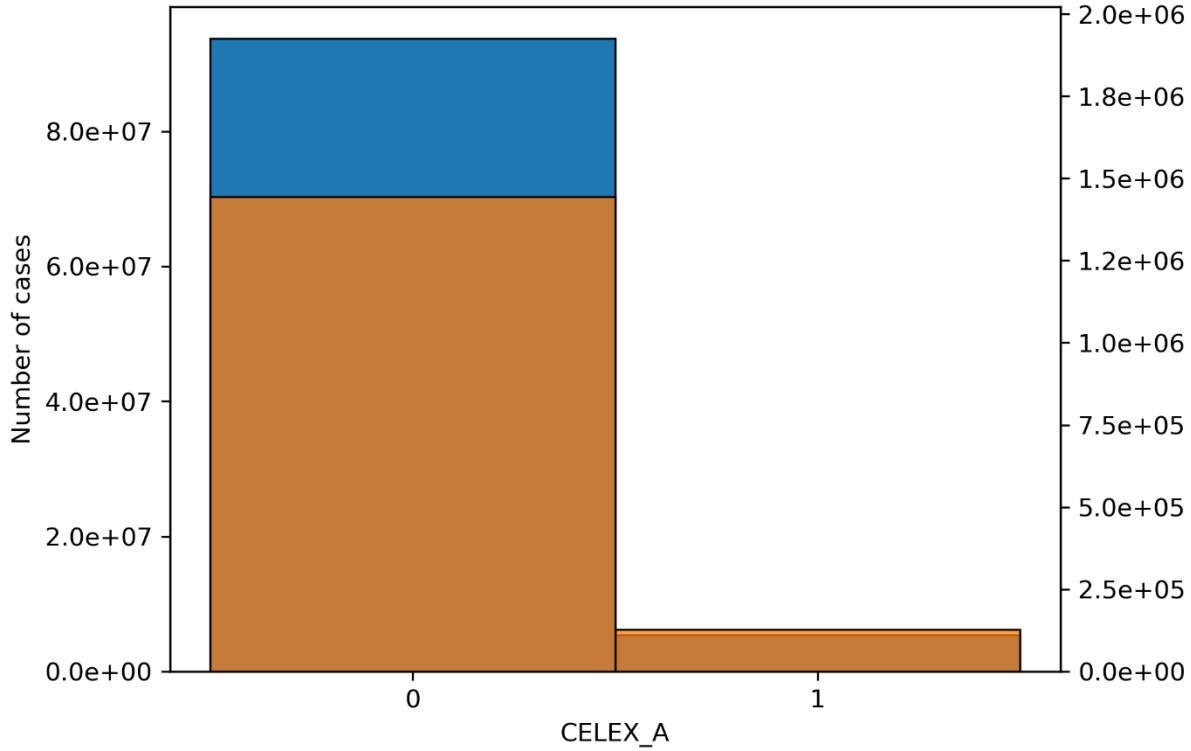
This appendix contains the results of the study per feature, including odds ratios, importances, and histogram plots. Each feature has one plot that contains two overlapping histograms: the first histogram, shaded blue, concerns the frequency of the values for a certain feature in the entire set of occurrences (i.e., deletions and non-deletions combined); the second, shaded in orange, concerns the frequency of the values for a certain feature in only the set of deletions. Because the latter histogram is overlaid on the former, the choice was made to make the orange histogram partly transparent. In places that the bars of the two histograms overlaps, this results in a brownish color. It should be noted that the two histograms make use of the same *x*-scale but differ in their *y*-scale: the *y*-scale for the first (blue) histogram can be found to the lefthand side of the plot, and the *y*-scale for the second (orange) histogram can be found to the righthand side of the plot. In order more clearly differentiate the two histograms from each other, the upper limit of the *y*-scale for the deletion histograms were increased proportionally by one third relative to its default value, meaning the bars for the orange deletion histograms are visually (in most cases) lower than those for the blue histograms. Lastly, for some plots, additional transformations were carried out (e.g., a logarithmic scale). In those cases, how the plot was altered is stated explicitly directly below the image.

CELEX_A

Statistics for CELEX_A

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.79499	0.00652	2.2144	0.000911	0.92998
2	0.74430	0.01056	2.1050	0.000421	0.85289
3	0.79462	0.00656	2.2136	0.000659	0.69219

Histogram of CELEX_A

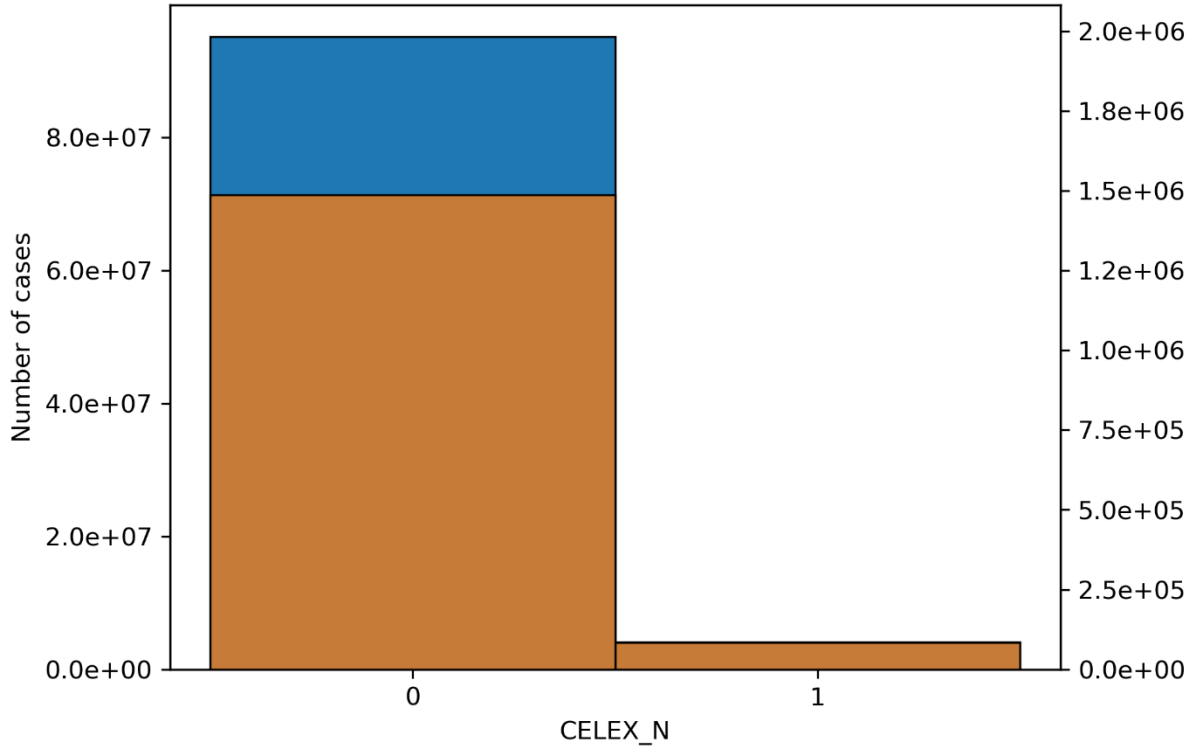


CELEX_N

Statistics for CELEX_N

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.166623	0.02584	1.1813	0.00019	0.69714
2	0.02461	0.27976	1.0249	7.81×10^{-5}	0.46342
3	0.16739	0.0257	1.1822	0.000261	0.9712

Histogram of CELEX_N

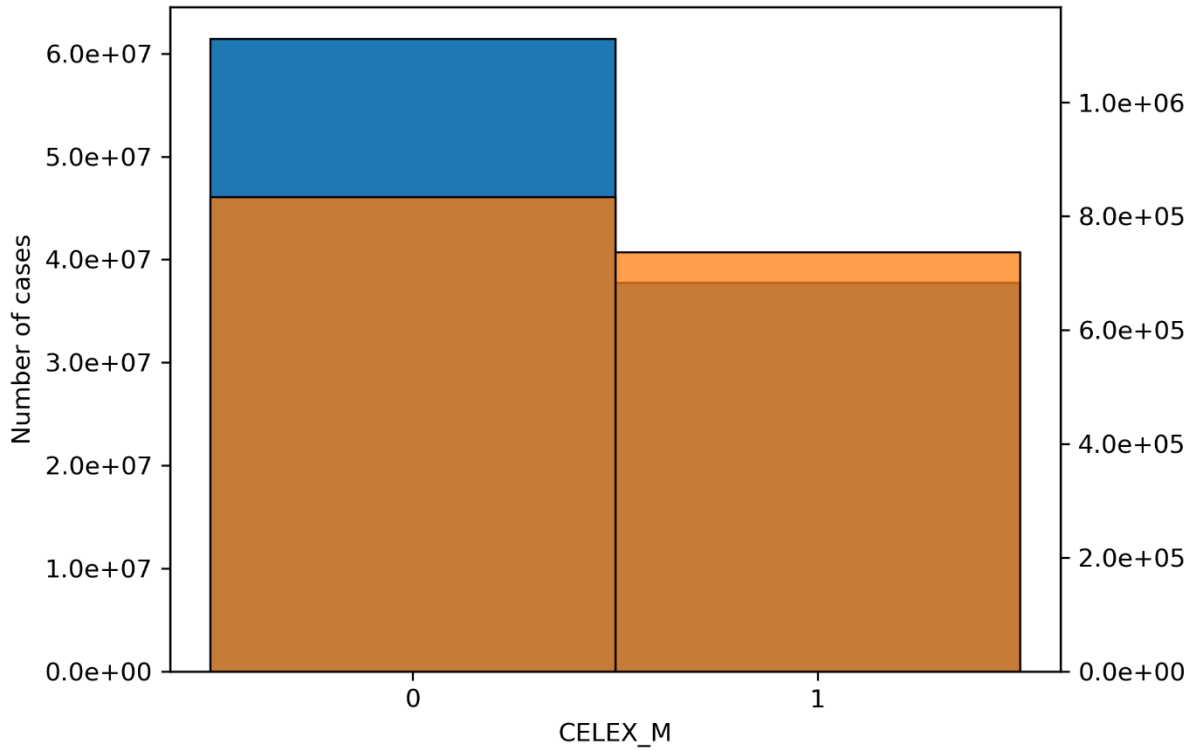


CELEX_M

Statistics for CELEX_M

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.083686	0.02281	1.0873	0.00019	0.63986
2	0.08844	0.02407	1.0925	7.81×10^{-5}	0.41376
3	0.08370	0.02273	1.0873	0.000261	0.5425

Histogram of CELEX_M

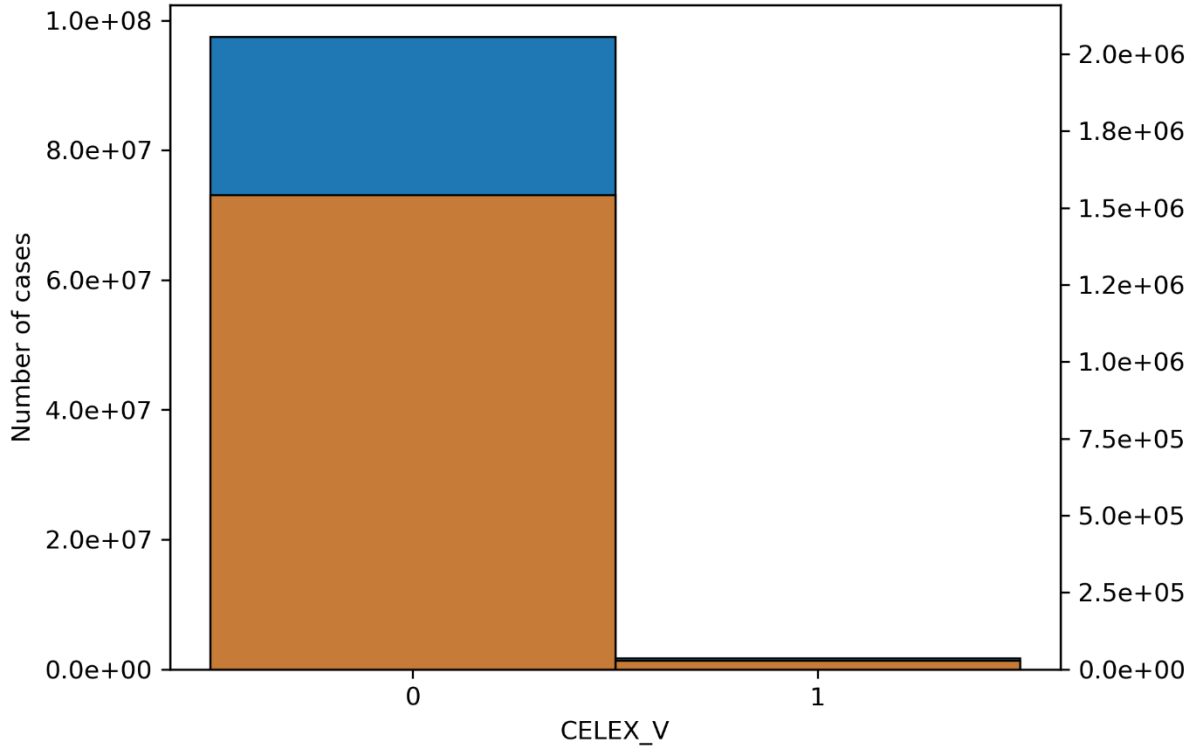


CELEX_V

Statistics for CELEX_V

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.639289	0.01538	1.8951	7.48×10^{-5}	0.5968
2	0.80516	0.01488	2.2371	6.56×10^{-5}	0.44174
3	0.63900	0.01523	1.8946	5.96×10^{-5}	0.70537

Histogram of CELEX_V

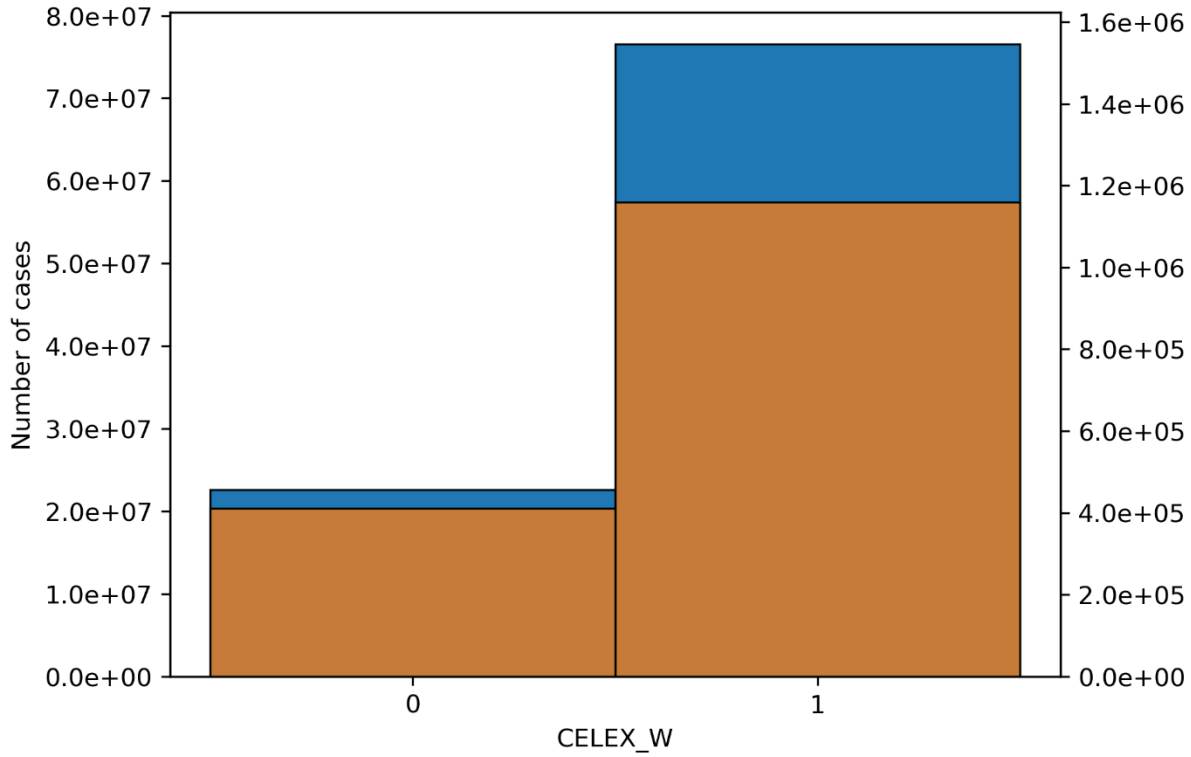


CELEX_W

Statistics for CELEX_W

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.672452	0.00438	1.9590	0.001094	0.48377
2	0.46247	0.00672	1.5880	0.000831	0.6917
3	0.67243	0.00434	1.9590	0.000733	0.39187

Histogram of CELEX_W

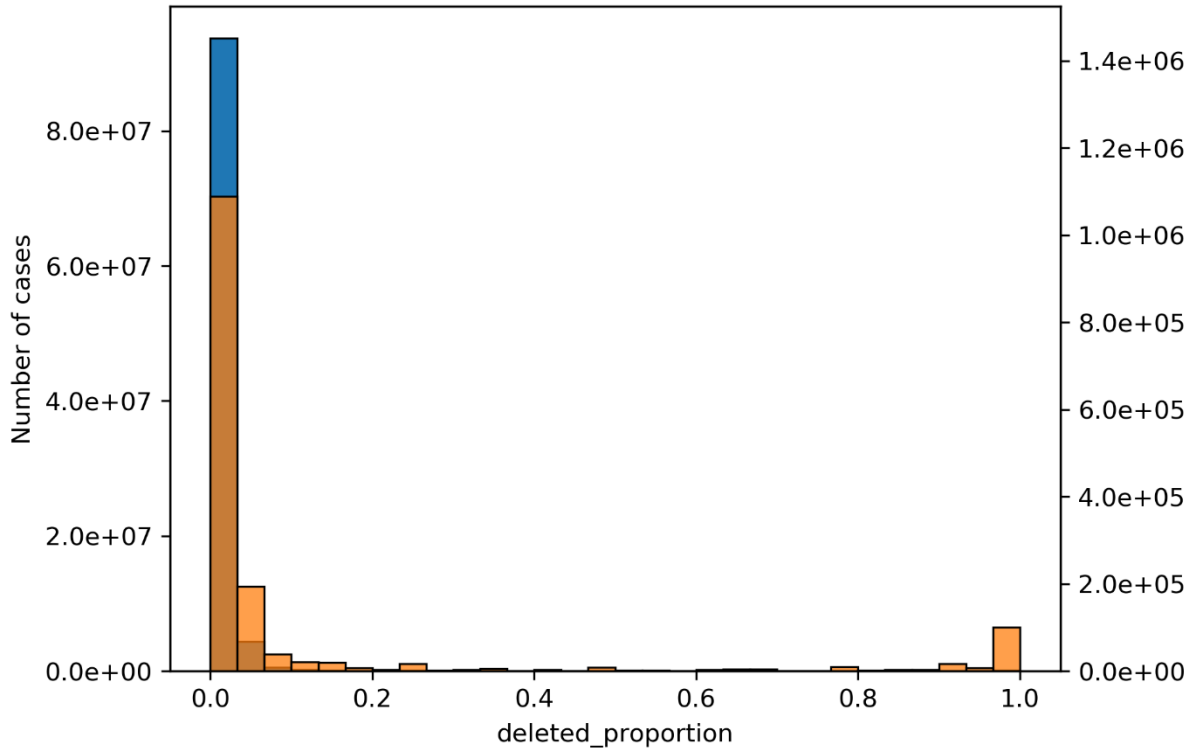


deleted_proportion

Statistics for deleted_proportion

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	25.22862	0.01027	9.05×10^{10}	0.117215	0.12346
2	20.46463	0.01161	7.72×10^{8}	0.089307	0.31199
3	25.23111	0.01024	9.07×10^{10}	0.113801	0.20004

Histogram of deleted_proportion



hour

Statistics for hour_0

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.107739	0.0565	1.1138	4.47×10^{-5}	0.80704
2	0.04896	0.15704	1.0502	9.27×10^{-5}	2.231
3	0.10772	0.05674	1.1137	9.38×10^{-5}	107389

Statistics for hour_1

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.260906	0.03566	1.2981	0.000227	1.22007
2	0.11615	0.04688	1.1232	9.56×10^{-5}	1.60533
3	0.26088	0.03615	1.2981	0.000102	0.71542

Statistics for hour_2

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.367222	0.02456	1.4437	3.58×10^{-5}	0.6817
2	0.15248	0.05012	1.1647	1.54×10^{-5}	0.70054
3	0.36733	0.02476	1.4439	7.11×10^{-5}	117472

Statistics for hour_3

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.4126	0.02022	1.5107	1.06×10^{-5}	0.82919
2	0.19222	0.05525	1.2119	1.32×10^{-5}	1.30171
3	0.41236	0.02062	1.5104	3.46×10^{-5}	0.91614

Statistics for hour_4

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.299511	0.04425	1.3492	4.73×10^{-6}	0.70188
2	0.09961	0.11616	1.1047	2.76×10^{-6}	0.881
3	0.29922	0.04448	1.3488	5.34×10^{-6}	0.65606

Statistics for hour_5

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.054618	0.25457	1.0561	3.27×10^{-6}	0.83649
2	-0.02438	-0.55802	0.9759	2.27×10^{-6}	0.46454
3	0.05558	0.2542	1.0572	2.32×10^{-6}	0.56512

Statistics for hour_6

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.17454	-0.0701	0.8398	2.70×10^{-6}	0.84725
2	-0.09238	-0.06246	0.9118	2.71×10^{-6}	0.47914
3	-0.17491	-0.06998	0.8395	2.12×10^{-6}	0.47338

Statistics for hour_7

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.11464	-0.05576	0.8917	7.93×10^{-6}	0.46775
2	-0.01861	-0.28795	0.9816	4.24×10^{-6}	0.55807
3	-0.11498	-0.05519	0.8914	7.50×10^{-6}	0.67629

Statistics for hour_8

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.41077	-0.01262	0.6631	1.01×10^{-5}	0.8773
2	-0.24403	-0.00911	0.7835	5.19×10^{-6}	0.72016
3	-0.41084	-0.01258	0.6631	9.76×10^{-6}	0.55989

Statistics for hour_9

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.52654	-0.00833	0.5906	6.87×10^{-5}	1.40294
2	-0.33334	-0.00912	0.7165	1.99×10^{-5}	0.81447
3	-0.52649	-0.00836	0.5907	4.52×10^{-5}	0.82553

Statistics for hour_10

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.45506	-0.0064	0.6344	1.07×10^{-5}	0.70062
2	-0.30295	-0.00715	0.7386	1.45×10^{-5}	1.76574
3	-0.45507	-0.00646	0.6344	3.41×10^{-5}	1.08788

Statistics for hour_11

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.40407	-0.00826	0.6676	6.24×10^{-6}	0.60161
2	-0.26967	-0.01398	0.7636	5.46×10^{-6}	0.57007
3	-0.40398	-0.00824	0.6677	7.14×10^{-6}	0.86485

Statistics for hour_12

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.3451	-0.00801	0.7081	4.17×10^{-6}	0.47876
2	-0.23843	-0.02293	0.7879	2.72×10^{-6}	0.63739
3	-0.34508	-0.00817	0.7082	3.43×10^{-6}	0.32193

Statistics for hour_13

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.34689	-0.01444	0.7069	4.25×10^{-6}	0.43971
2	-0.22335	-0.01914	0.7998	3.10×10^{-6}	0.48822
3	-0.34706	-0.01466	0.7068	2.98×10^{-6}	0.52955

Statistics for hour_14

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.35391	-0.00814	0.7019	5.53×10^{-6}	0.47202
2	-0.22479	-0.00891	0.7987	3.11×10^{-6}	0.51541
3	-0.35400	-0.0083	0.7019	3.20×10^{-6}	0.59415

Statistics for hour_15

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.32648	-0.01007	0.7215	4.62×10^{-6}	0.39003
2	-0.20433	-0.01699	0.8152	2.52×10^{-6}	0.45471
3	-0.32629	-0.01028	0.7216	3.81×10^{-6}	0.44423

Statistics for hour_16

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.28989	-0.01181	0.7483	2.92×10^{-6}	0.419
2	-0.17774	-0.01229	0.8372	2.45×10^{-6}	0.40105
3	-0.29009	-0.0121	0.7482	3.78×10^{-6}	0.33729

Statistics for hour_17

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.23764	-0.01037	0.7885	3.71×10^{-6}	0.39529
2	-0.13401	-0.02448	0.8746	2.84×10^{-6}	0.33189
3	-0.23799	-0.01042	0.7882	2.75×10^{-6}	0.51708

Statistics for hour_18

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.18075	-0.02331	0.8346	4.06×10^{-6}	0.55364
2	-0.09351	-0.05465	0.9107	3.29×10^{-6}	0.30522
3	-0.18092	-0.02356	0.8345	3.12×10^{-6}	0.8935

Statistics for hour_19

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.20797	-0.01908	0.8122	2.97×10^{-6}	0.24178
2	-0.09232	-0.02881	0.9118	2.94×10^{-6}	0.26818
3	-0.20799	-0.01941	0.8122	3.97×10^{-6}	0.52809

Statistics for hour_20

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.21098	-0.00915	0.8098	7.55×10^{-6}	0.41492
2	-0.07671	-0.05364	0.9262	3.94×10^{-6}	0.81257
3	-0.21112	-0.00922	0.8097	5.04×10^{-6}	0.34348

Statistics for hour_21

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.15396	-0.0232	0.8573	5.51×10^{-6}	0.71783
2	-0.03846	-0.09563	0.9623	3.49×10^{-6}	0.3906
3	-0.15381	-0.02327	0.8574	3.11×10^{-6}	0.35042

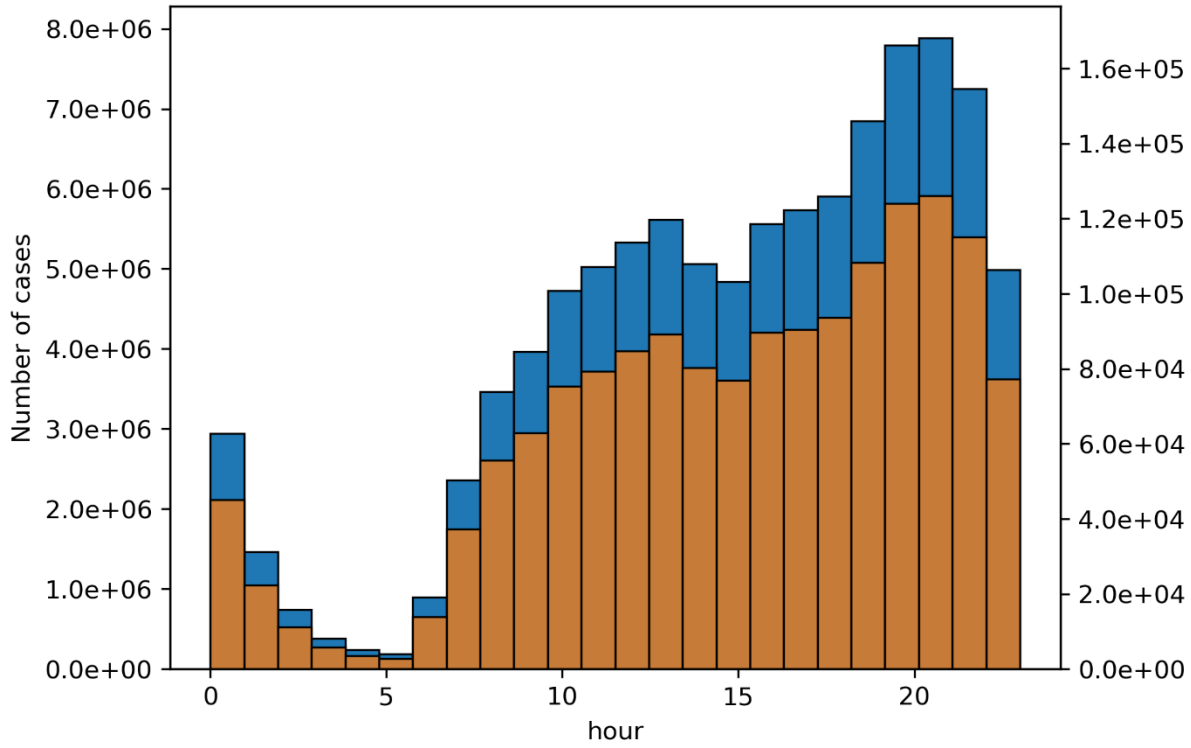
Statistics for hour_22

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.07278	-0.02613	0.9298	1.01×10^{-5}	0.54723
2	-0.01958	-0.19288	0.9806	4.63×10^{-6}	0.20681
3	-0.07320	-0.02561	0.9294	8.67×10^{-6}	0.51064

Statistics for hour_23

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	NA	NA	NA	2.08×10^{-5}	0.914
2	NA	NA	NA	1.00×10^{-5}	0.84837
3	NA	NA	NA	4.51×10^{-5}	1466

Histogram of hour

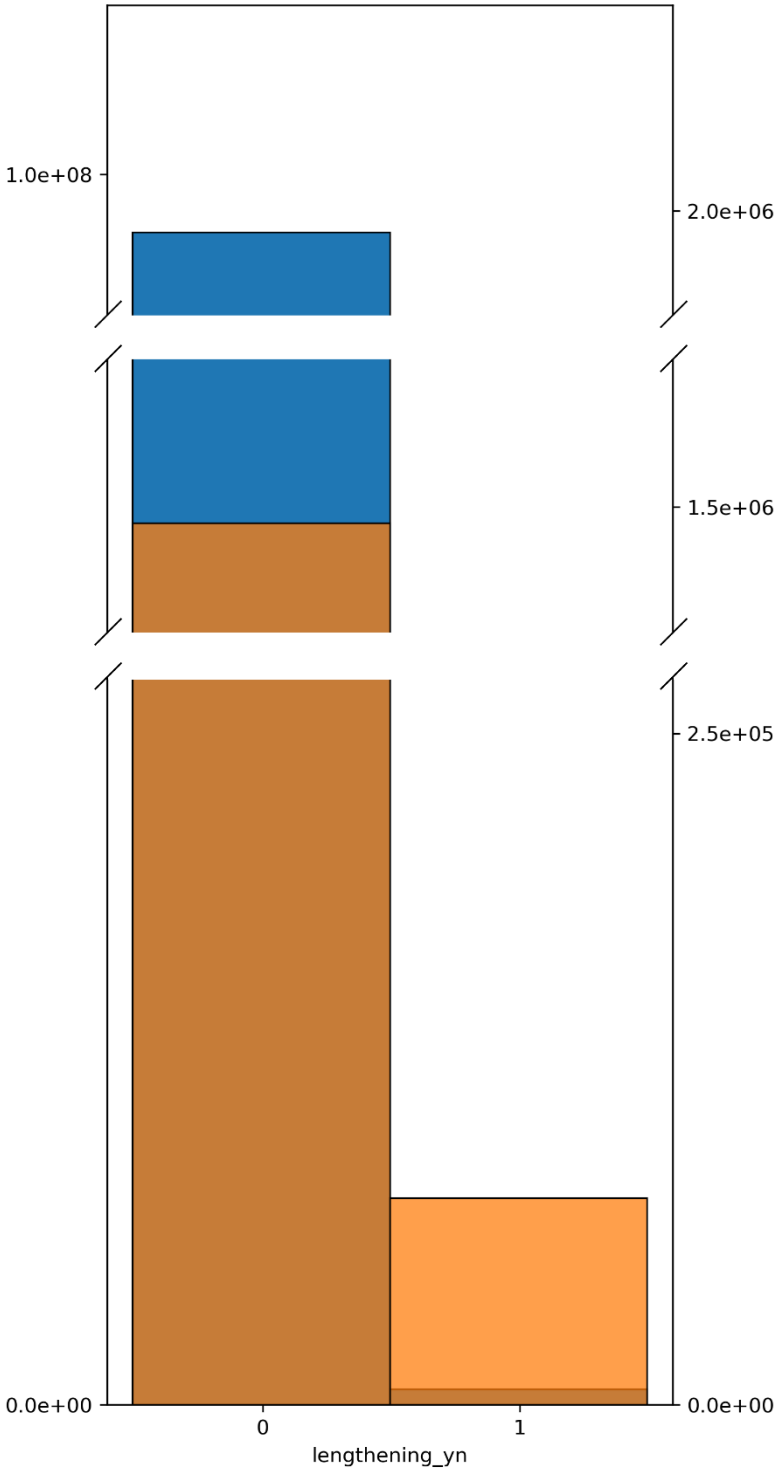


lengthening_yn

Statistics for lengthening_yn

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	1.689897	0.00683	5.4189	0.010888	0.47808
2	1.84933	0.0053	6.3556	0.006522	0.35149
3	1.69000	0.00674	5.4195	0.008731	0.44006

Histogram of lengthening_yn



Note that the y-axes of the histogram have been broken twice in order to allow the low frequency of *lengthening* = 1 for the blue histogram to remain visible.

righthand

Statistics for righthand_consonant

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.003459	0.36743	1.0035	0.002216	0.33489
2	0.05477	0.03143	1.0563	0.00187	0.58309
3	0.00342	0.37245	1.0034	0.004166	0.64685

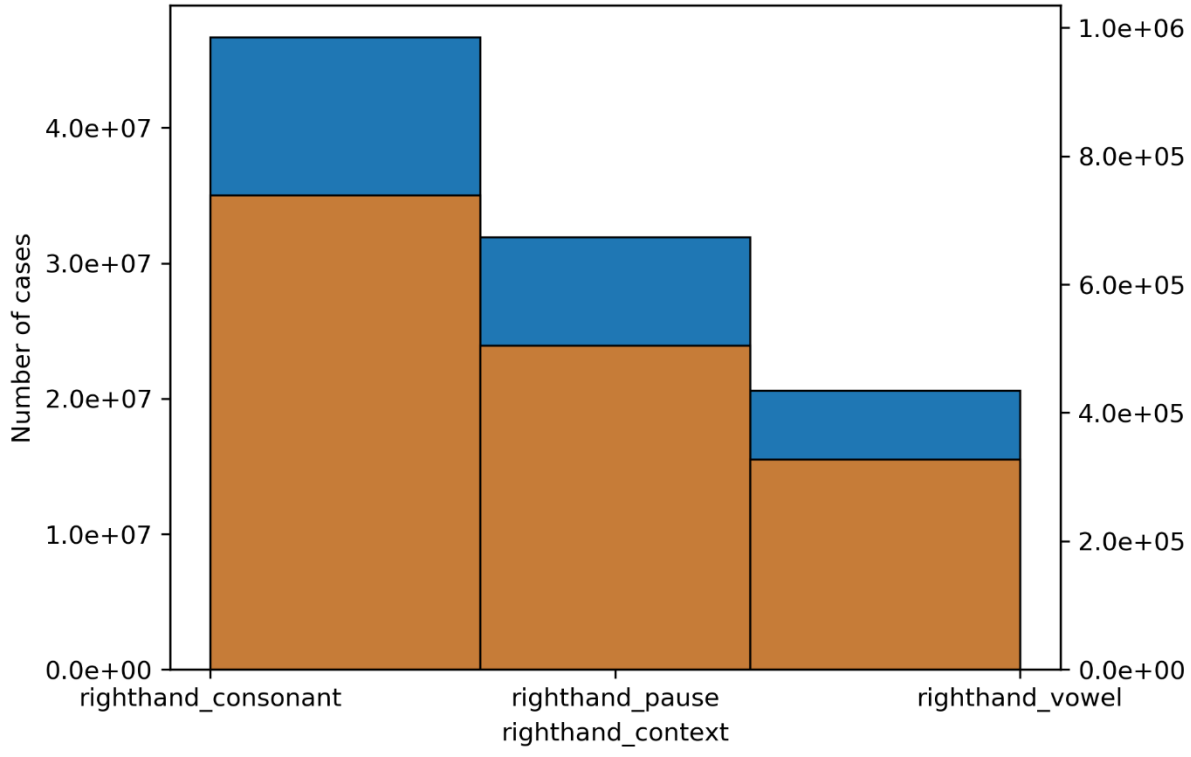
Statistics for righthand_pause

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.931115	0.00398	2.5373	0.011971	0.18131
2	1.05402	0.00376	2.8692	0.005864	0.27486
3	0.93094	0.00396	2.5369	0.014352	0.43791

Statistics for righthand_vowel

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	NA	NA	NA	0.001301	0.55488
2	NA	NA	NA	0.000872	0.86761
3	NA	NA	NA	0.00167	112843

Histogram of righthand

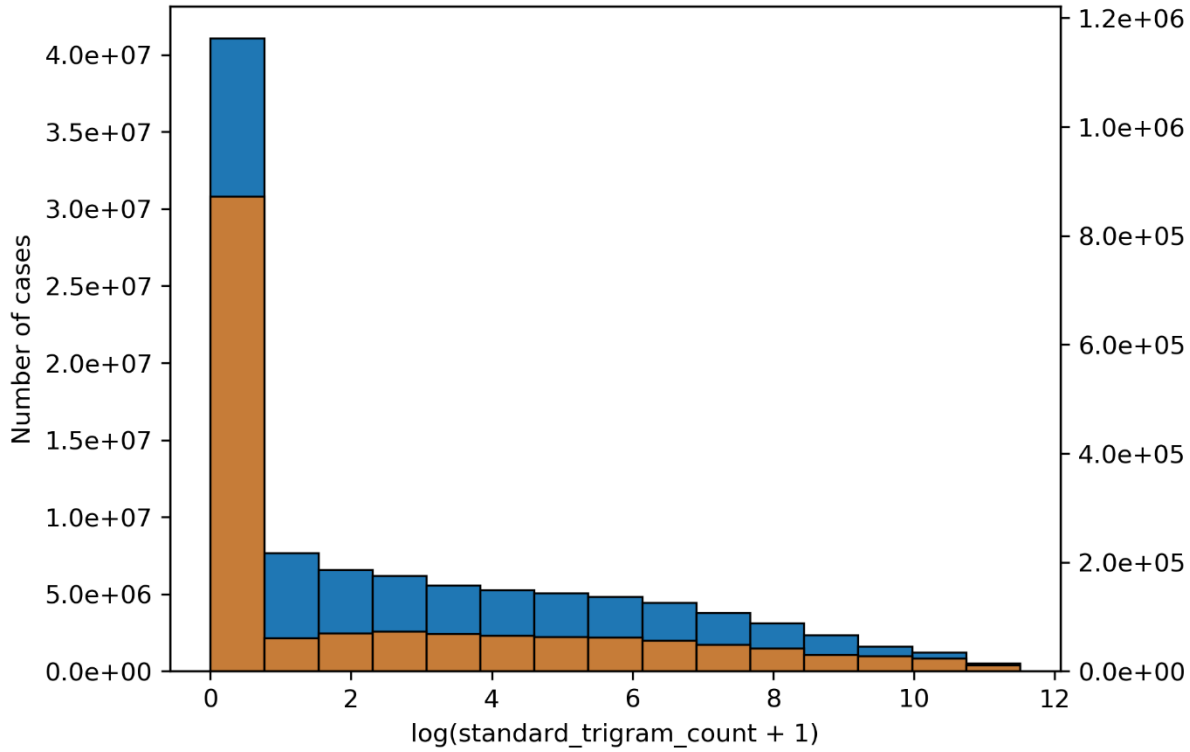


standard_trigram_count_log

Statistics for standard_trigram_count_log

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-2.79599	-0.00175	0.0611	0.378005	0.06633
2	-3.19407	-0.00161	0.0410	0.251139	0.16183
3	-2.79578	-0.00176	0.0611	0.370387	0.0944

Histogram of standard_tirgram_count



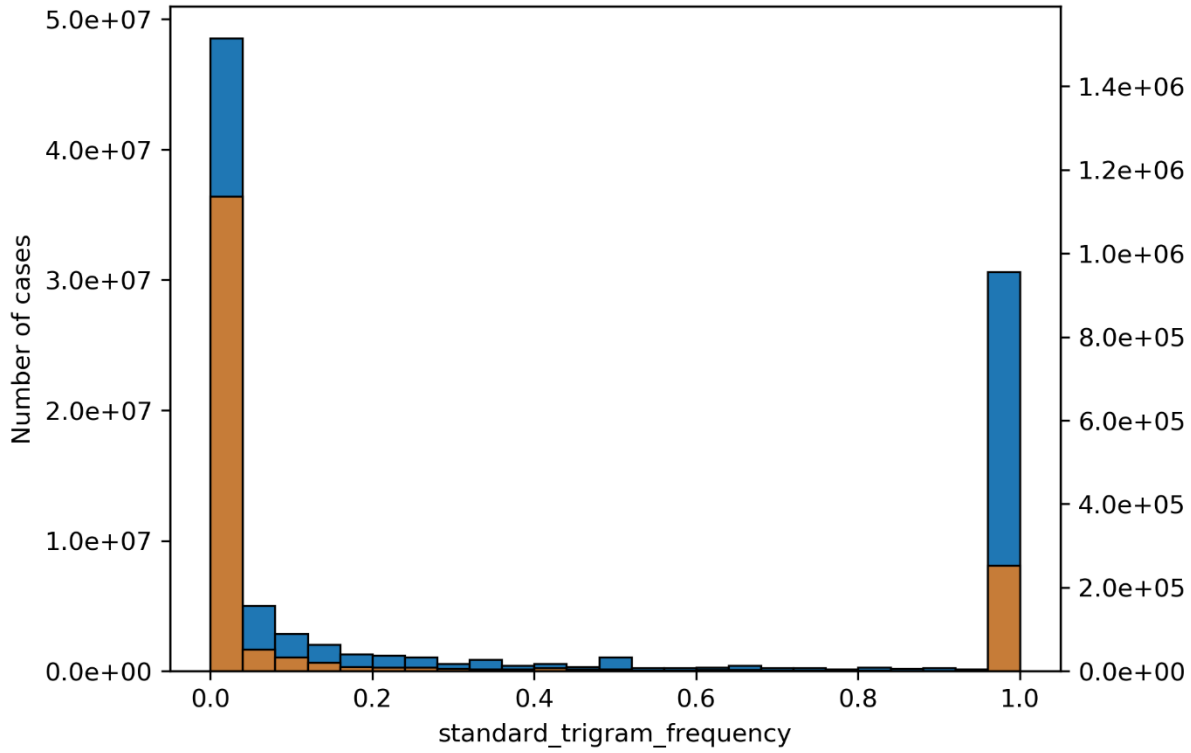
Note that the above graph represents values of *standard_trigram_count* before a log was applied. These values were then fed through the function ln(x+1) to yield a graph that not sensitive to outliers and that does not have values below zero.

standard_trigram_frequency

Statistics for standard_trigram_frequency

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-2.38678	-0.00147	0.0919	0.139049	0.09957
2	-2.21148	-0.00168	0.1095	0.111933	0.16815
3	-2.38667	-0.00149	0.0919	0.15023	0.21317

Histogram of standard_trigram_frequency

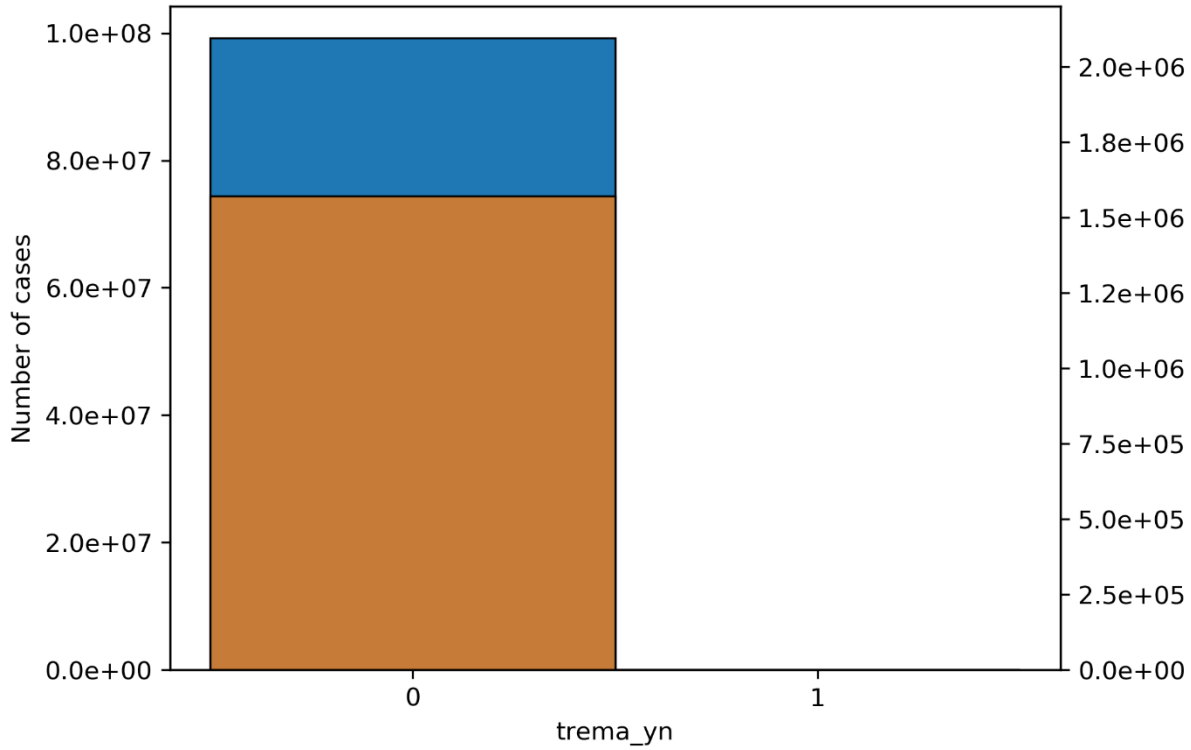


trema_yn

Statistics for trema_yn

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-12.0795	-0.02636	0.000	3.67×10^{-5}	0.8002
2	-1625695	-0.03139	0.000	1.83×10^{-5}	0.38524
3	-1196556	-0.00193	0.000	3.10×10^{-5}	0.58843

Histogram of trema_yn

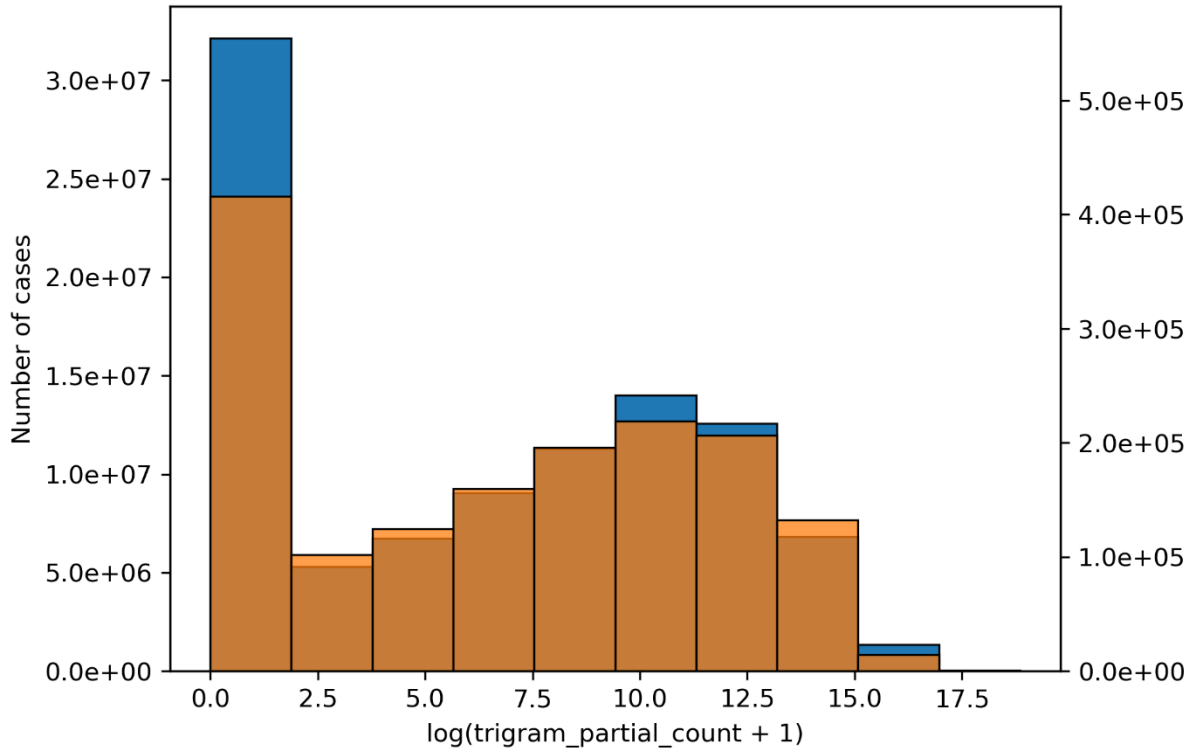


trigram_partial_count

Statistics for trigram_partial_count

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.64154	-0.01184	0.5265	0.150242	0.14132
2	-0.41496	-0.02383	0.6604	0.083449	0.32851
3	-0.64153	-0.01157	0.5265	0.154183	0.18191

Histogram of trigram_partial_count



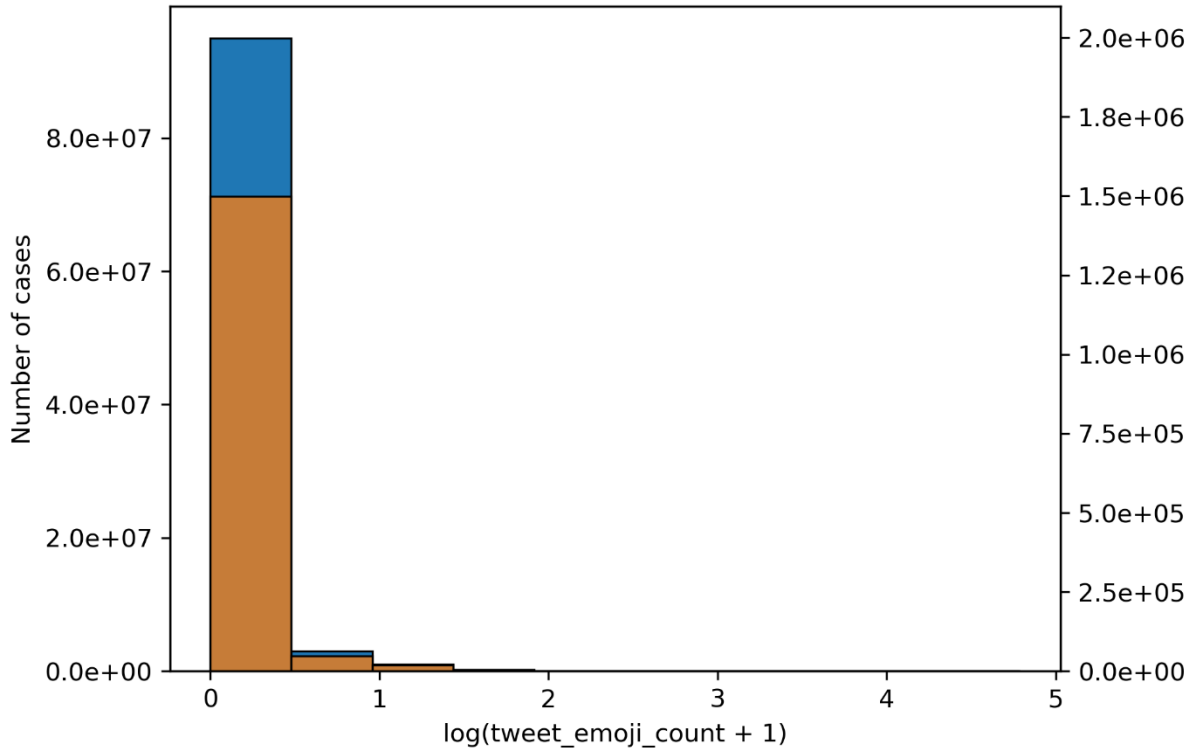
Note that the above graph represents *trigram_partial_count* fed through the function ln(x+1) to yield a graph that not sensitive to outliers and that does not have values below zero.

tweet_emoji_count

Statistics for tweet_emoji_count

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	18.91182	0.04018	1.64×10^{8}	0.000493	0.37429
2	2.76596	0.08296	15.8943	0.000202	0.39865
3	18.89000	0.04017	1.60×10^{8}	0.000482	0.47529

Histogram of tweet_emoji_count



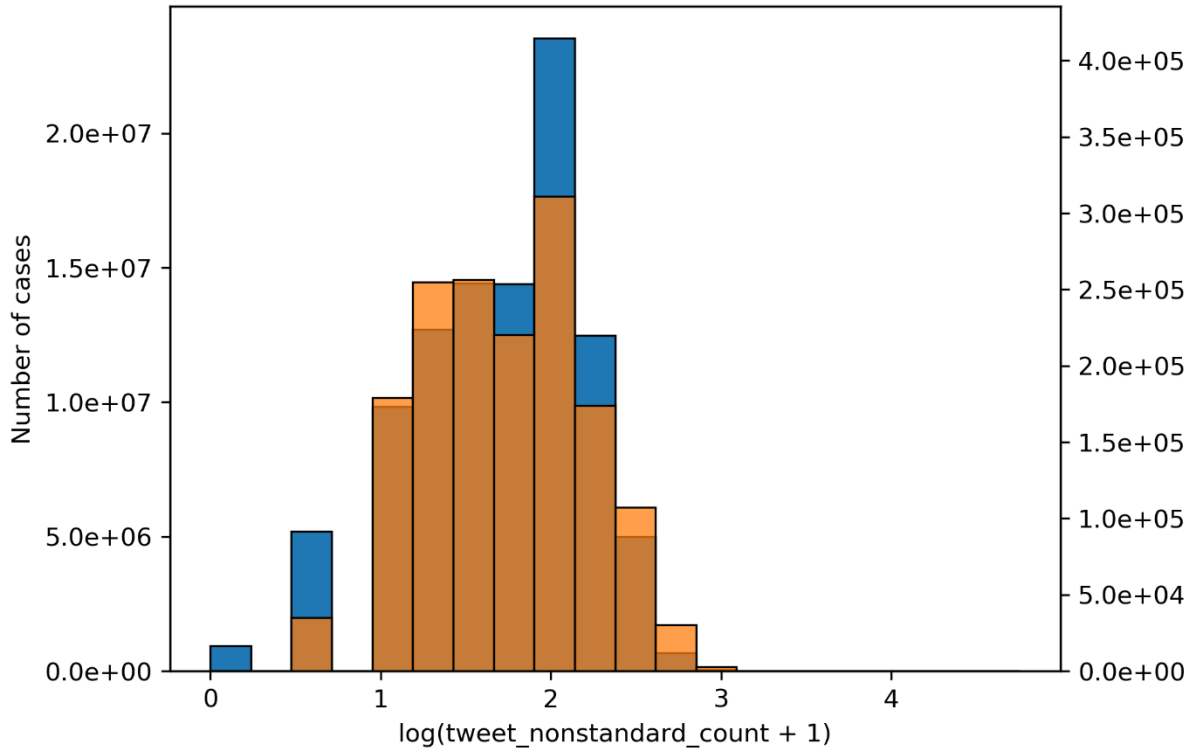
Note that the above graph represents *tweet_emoji_count* before a log was applied fed through the function ln(x+1) to yield a graph that not sensitive to outliers and that does not have values below zero.

tweet_nonstandard_count

Statistics for tweet_nonstandard_count

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	15.66818	0.00619	6.38×10^{6}	0.014458	0.20975
2	18.79374	0.00365	1.45×10^{8}	0.010851	0.12796
3	15.67333	0.00624	6.41×10^{6}	0.013427	0.26961

Histogram of tweet_nonstandard_count



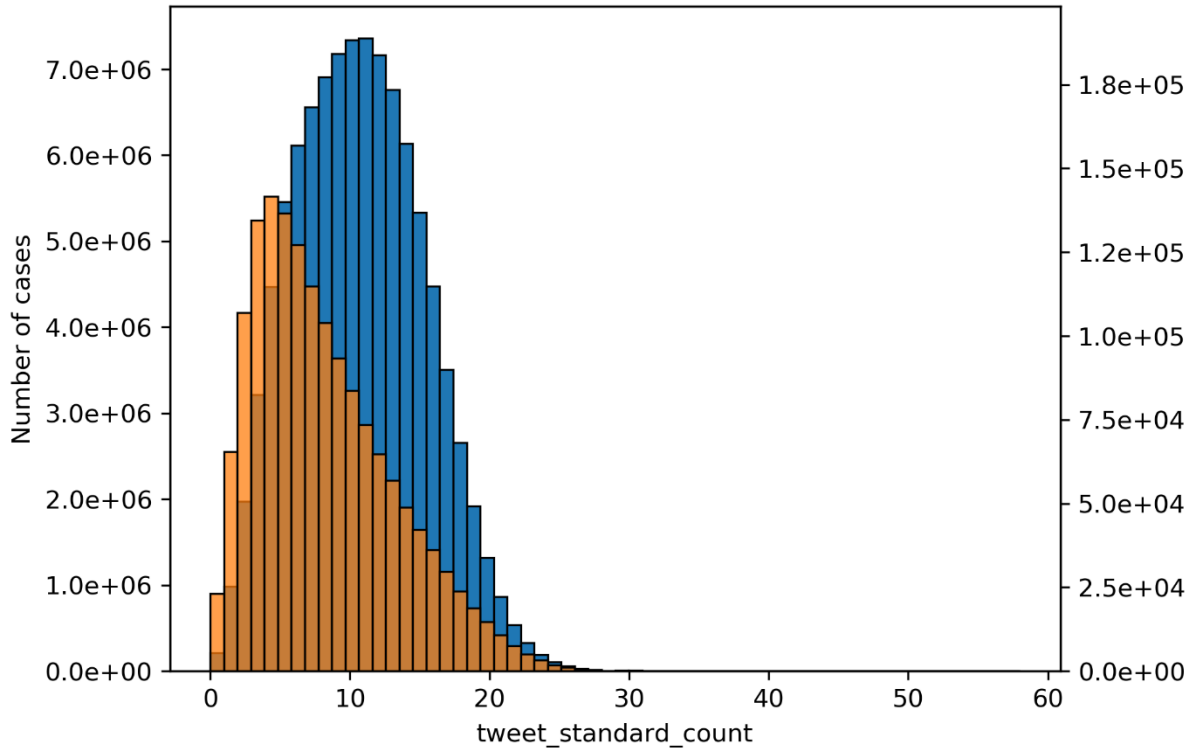
Note that the above graph represents *standard_trigram_count* fed through the function $\ln(x+1)$ to yield a graph that not sensitive to outliers and that does not have values below zero.

tweet_standard_count

Statistics for tweet_standard_count

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-5.78574	-0.00295	0.0031	0.062083	0.18835
2	-4.59035	-0.00292	0.0101	0.032518	0.47451
3	-5.78633	-0.00293	0.0031	0.056514	0.2746

Histogram of tweet_standard_count



user_gender

Statistics for user_gender_Mcertain

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.20163	-0.05163	0.8174	0.001598	0.61452
2	-0.06459	-0.10259	0.9375	0.000882	0.88652
3	-0.20408	-0.05018	0.8154	0.001758	0.51268

Statistics for user_gender_Mmaybe

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.268088	0.07637	1.3075	7.71×10^{-5}	0.48799
2	0.14816	0.09232	1.1597	2.26×10^{-5}	0.73982
3	0.26366	0.07492	1.3017	6.97×10^{-5}	0.39615

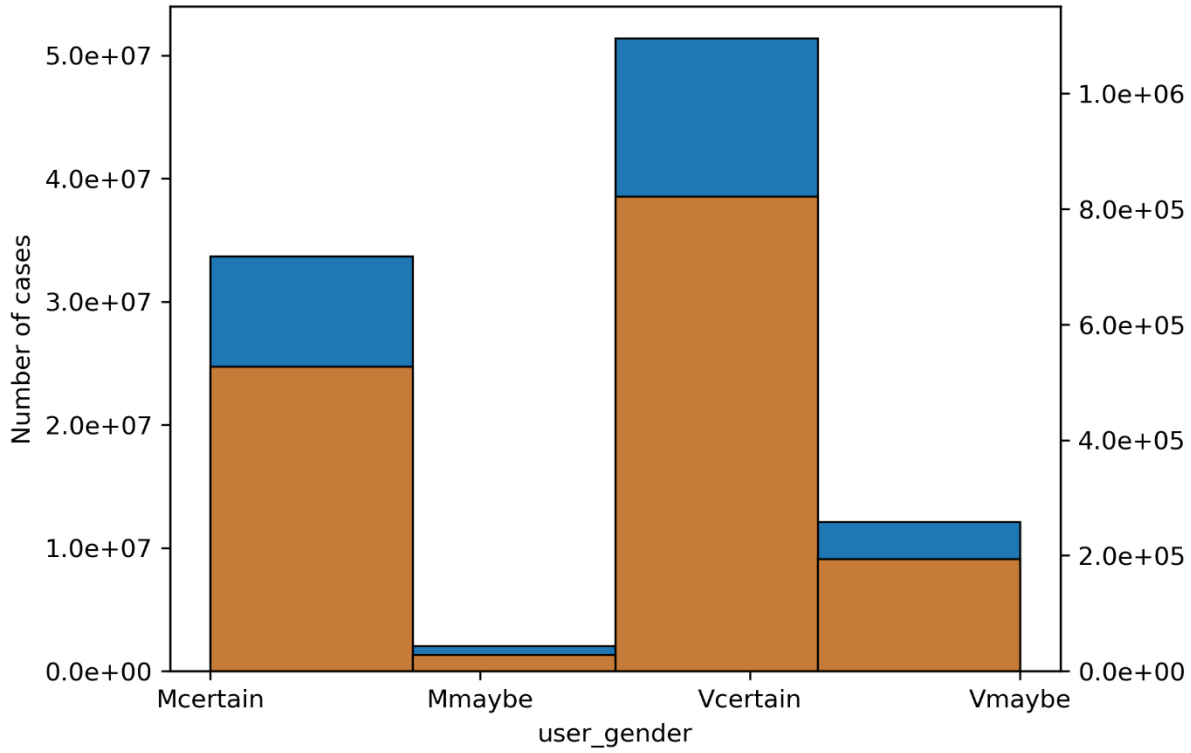
Statistics for user_gender_Vcertain

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.108567	0.12035	1.1147	0.000349	0.73554
2	0.24242	0.03012	1.2743	0.000131	0.68929
3	0.10640	0.12104	1.1123	0.000378	0.69675

Statistics for user_gender_Vmaybe

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.431101	0.04621	1.5390	0.000772	0.66743
2	0.38248	0.03904	1.4659	0.0003	109735
3	0.42806	0.04604	1.5343	0.000418	0.84284

Histogram of user_gender

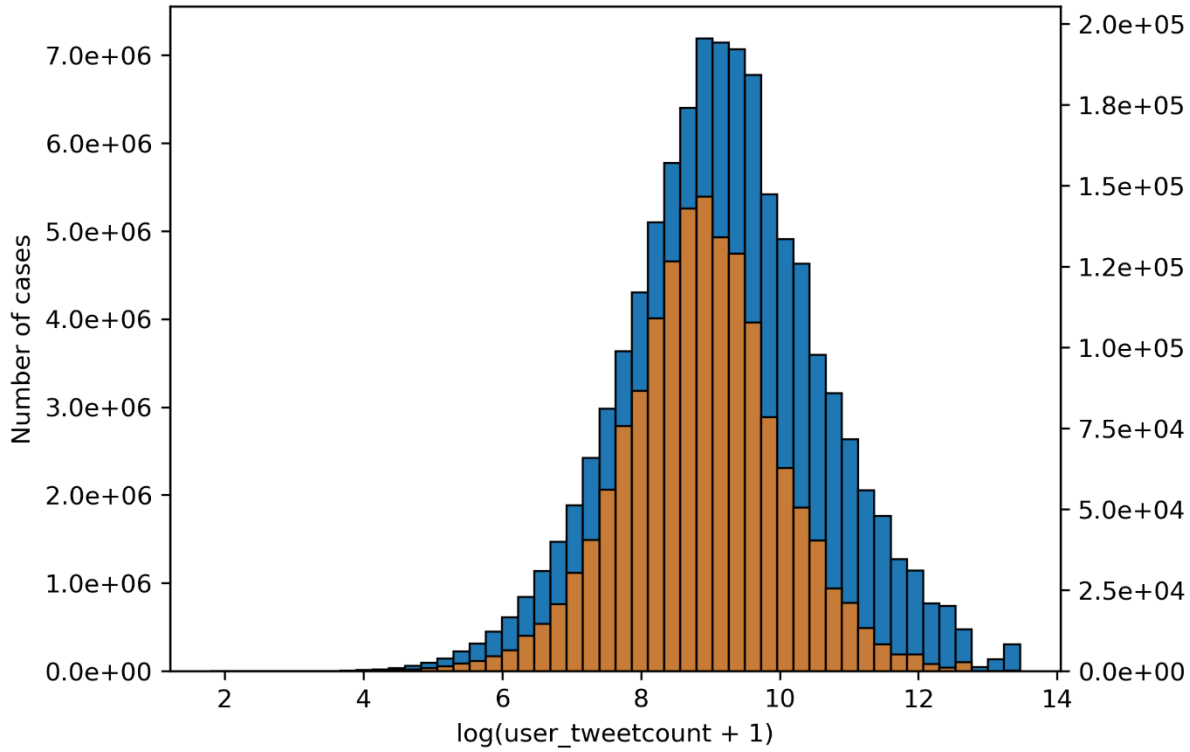


user_tweetcount_log

Statistics for user_tweetcount_log

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-2.51853	-0.02728	0.0806	0.021916	0.33844
2	-0.23867	-0.06123	0.7877	0.011545	0.2735
3	-2.51356	-0.02723	0.0810	0.020528	0.27456

Histogram of user_tweetcount



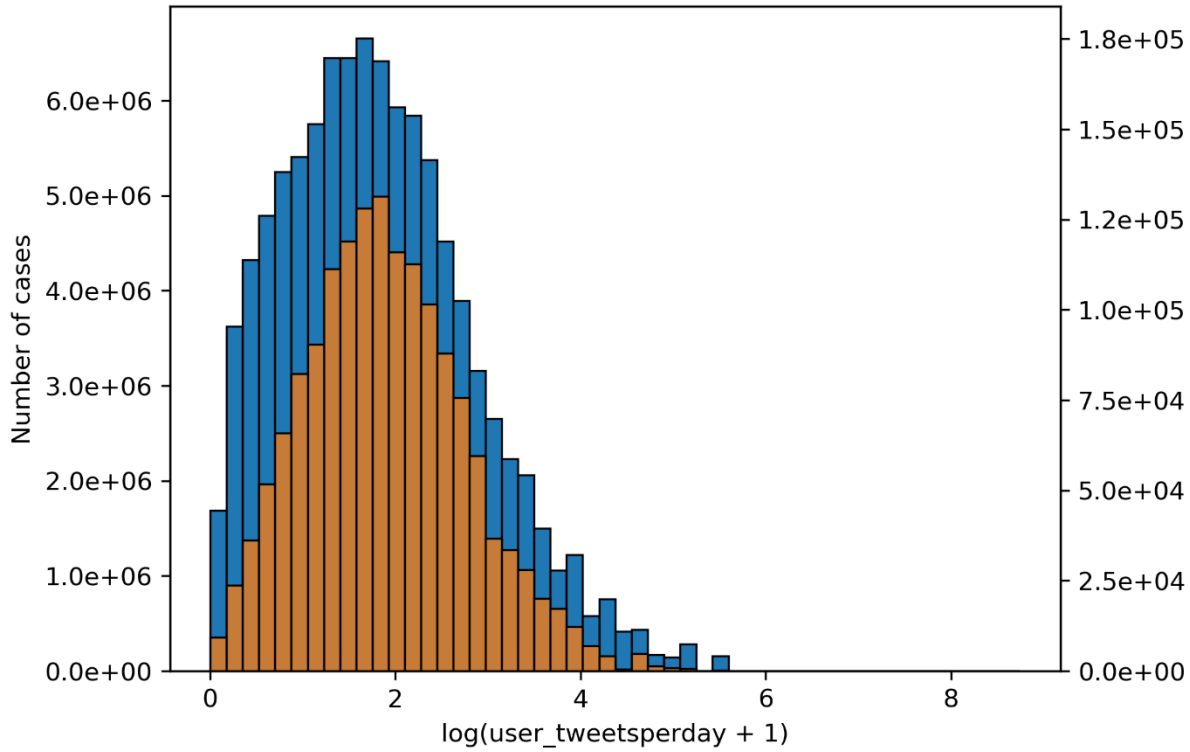
Note that the above graph represents *user_tweetcount* before a log was applied, but then fed through the function ln(x+1) to yield a graph that not sensitive to outliers and that does not have values below zero.

user_tweetsperday

Statistics for user_tweetsperday

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	24.14078	0.1229	3.05×10^{10}	0.005694	0.46678
2	0.88113	1.28198	2.4136	0.004254	0.51998
3	23.24333	0.12288	1.24×10^{10}	0.010166	0.44342

Histogram of user_tweetsperday



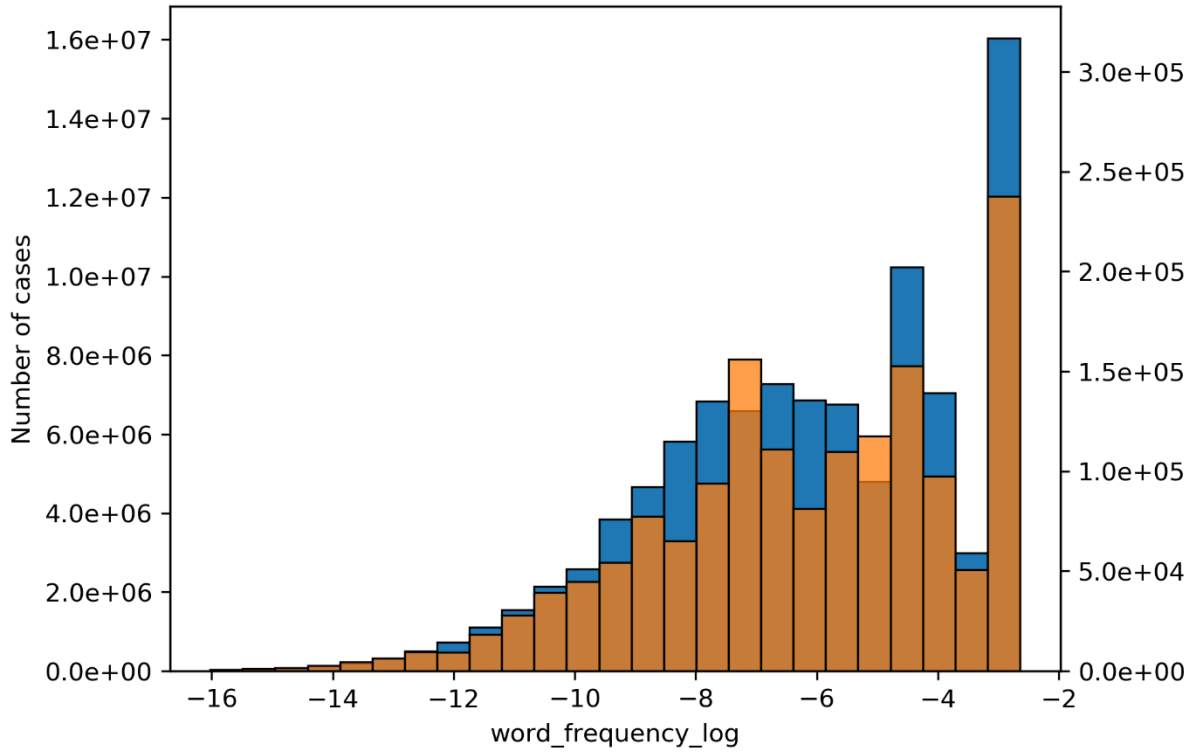
Note that the above graph represents *user_tweetsperday* fed through the function ln(x+1) to yield a graph that not sensitive to outliers and that does not have values below zero.

word_frequency_log

Statistics for word_frequency_log

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.228711	0.04309	1.2570	0.007608	0.22584
2	0.18828	0.05434	1.2072	0.008907	0.43149
3	0.22858	0.04304	1.2568	0.011264	0.5264

Histogram of word_frequency_log



year

Statistics for year_2010

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	2.186859	0.01273	8.9072	1.31×10^{-5}	2.55055
2	1.83440	0.01155	6.2614	1.52×10^{-6}	0.54233
3	2.18856	0.01264	8.9224	8.43×10^{-7}	0.74552

Statistics for year_2011

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	2.692569	0.00557	14.7696	0.005663	0.6717
2	1.96940	0.00822	7.1664	0.003347	0.68397
3	2.69289	0.00549	14.7743	0.005252	0.34273

Statistics for year_2012

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	2.759282	0.00586	15.7885	0.01341	0.44945
2	1.84423	0.00941	6.3232	0.006468	0.6378
3	2.76033	0.00591	15.8051	0.011039	0.32847

Statistics for year_2013

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	2.477413	0.00603	11.9104	0.005661	0.38289
2	1.55445	0.01169	4.7325	0.003558	0.56917
3	2.47822	0.00605	11.9200	0.006852	0.24303

Statistics for year_2014

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	1.694651	0.01335	5.4447	0.001006	0.73382
2	1.09796	0.01834	2.9980	0.000283	1.03518
3	1.69422	0.01341	5.4424	0.000596	0.88746

Statistics for year_2015

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.953185	0.02702	2.5940	0.006864	0.46761
2	0.68947	0.03257	1.9927	0.003881	0.64151
3	0.95219	0.02695	2.5914	0.006834	0.45814

Statistics for year_2016

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.53752	0.047	1.7118	0.008167	0.51949
2	0.34984	0.07216	1.4188	0.003545	0.76618
3	0.53620	0.04699	1.7095	0.008747	0.37239

Statistics for year_2017

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	0.248404	0.09891	1.2820	0.010603	0.29754
2	0.11690	0.17063	1.1240	0.003648	0.56714
3	0.24683	0.09919	1.2800	0.009916	0.3306

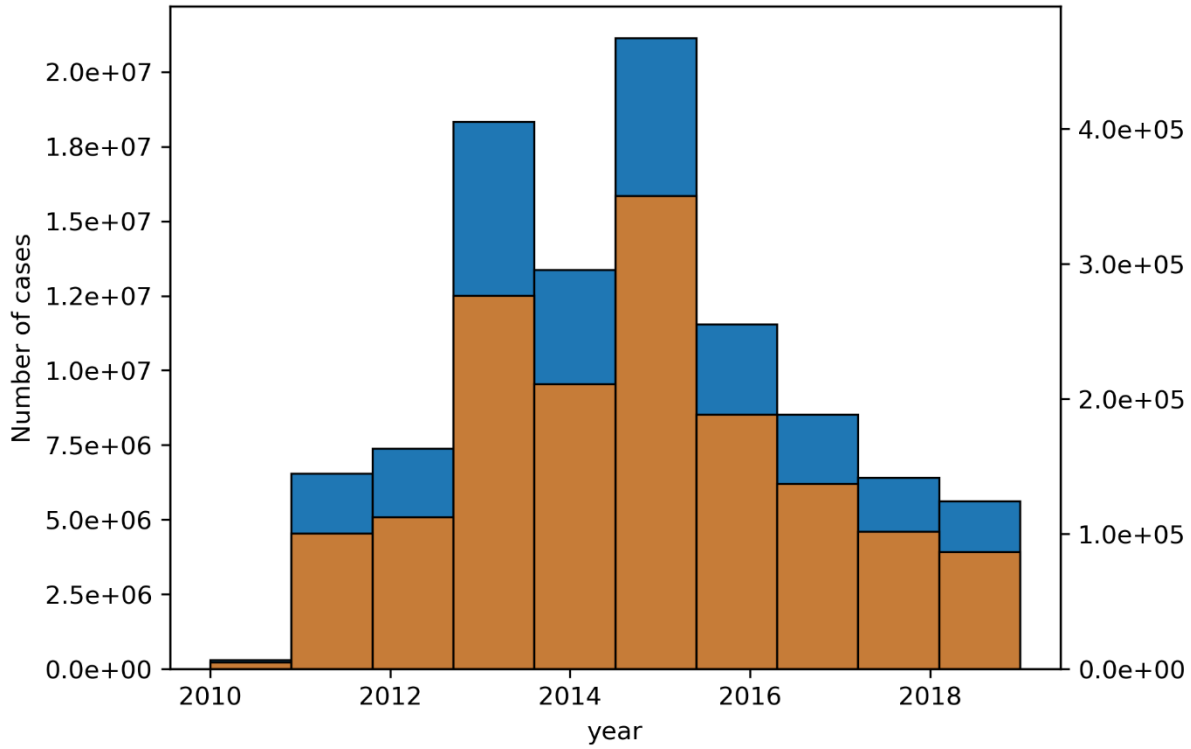
Statistics for year_2018

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	-0.01488	-1.37016	0.9852	0.00986	0.49397
2	-0.10477	-0.11256	0.9005	0.005208	0.53669
3	-0.01632	-1.24187	0.9838	0.009098	0.44492

Statistics for year_2019

Metric / *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	NA	NA	NA	0.008412	0.55018
2	NA	NA	NA	0.003159	0.62443
3	NA	NA	NA	0.005682	0.43766

Histogram of year

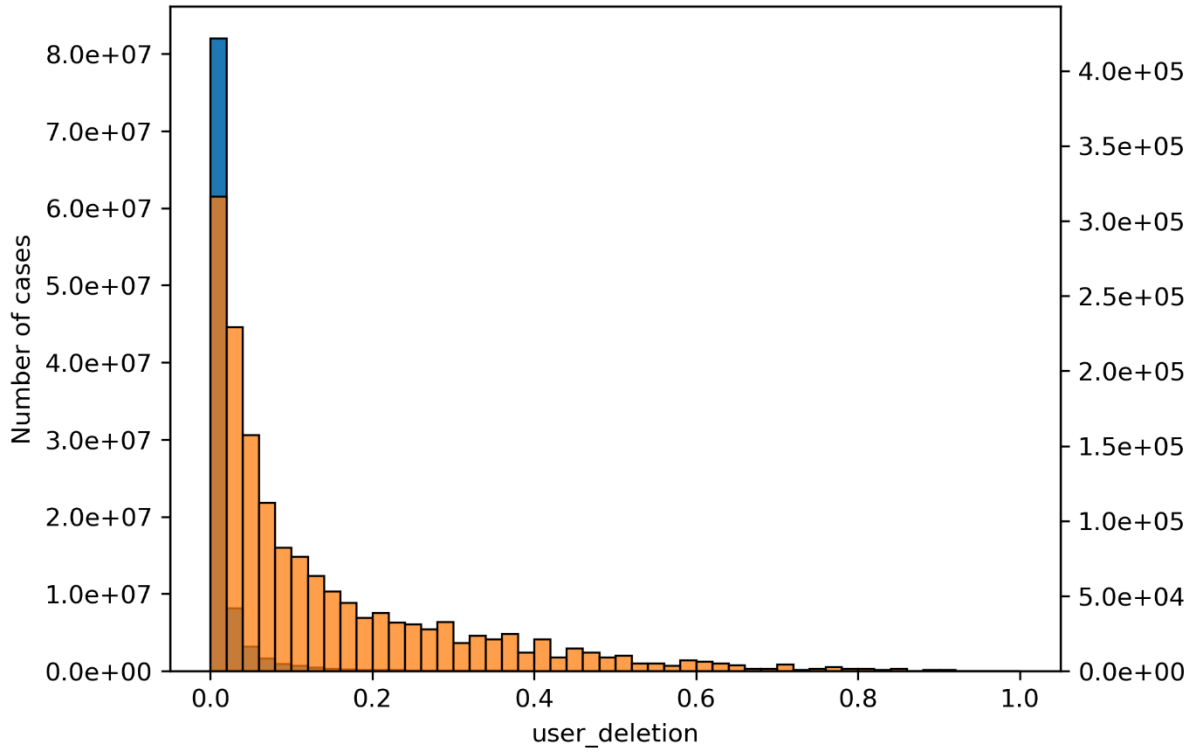


user_deletion

Statistics for user_deletion

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1					
2	24.59394	0.00739	4.80×10^{10}	0.3437	0.22336
3					

Histogram of user_deletion

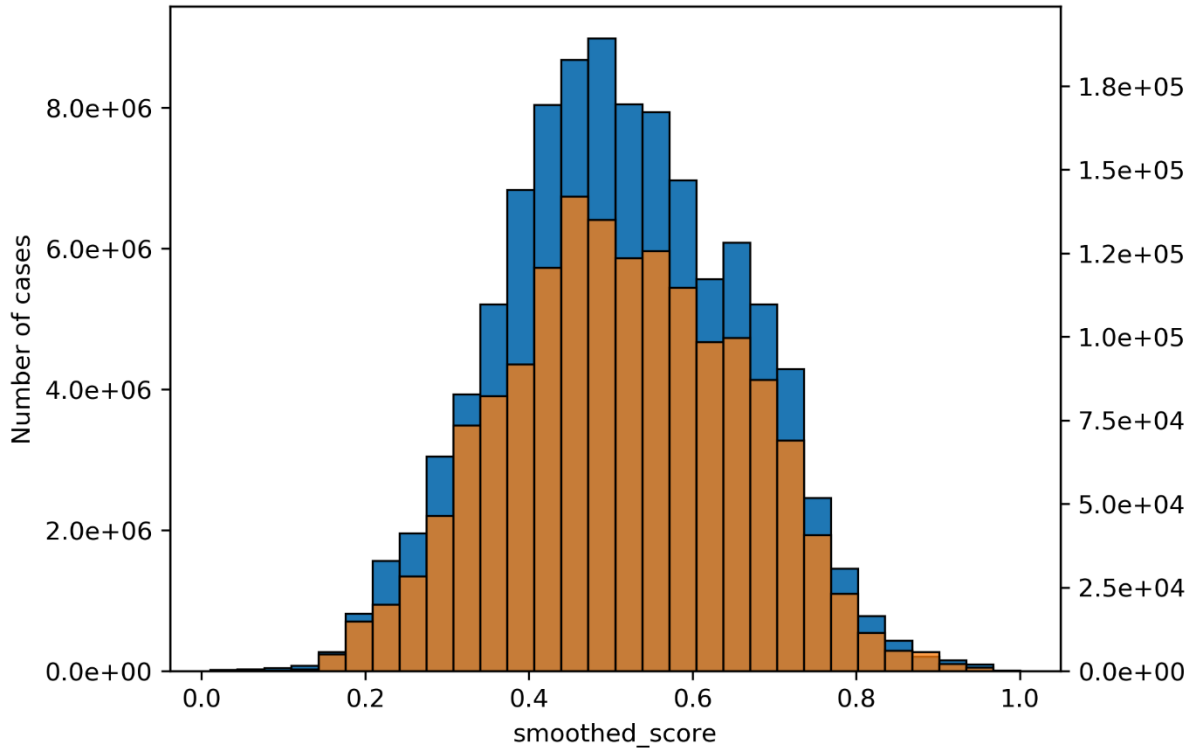


user_smoothed

Statistics for user_smoothed

Metric *Instance*	*Log odds (regression)*	*Coef. of var. (regression)*	*Odds ratio (regression)*	*Importance (ran. for.)*	*Coef of v. (ran. for.)*
1	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA
3	0.28457	0.14691	1.3292	0.0001700	0.38114

Histogram of user_smoothed



Appendix 3: EACH Approval

Ethics Assessment Committee Humanities

Faculty of Arts and Faculty of Philosophy, Theology and Religious Studies

Visiting Address
Erasmusplein 1
6525 HT Nijmegen

Postal Address
Postbus 9103
6500 HD Nijmegen

Radboud University
Faculty of Arts
Attn dr. B.J.M. van Halteren
Erasmusplein 1
6525 HT NIJMEGEN

Date	*Our Reference*	*Contact details*
23 February 2022	22U.003656	T: +31 (0)24 361 58 14 E: etc-gw@ru.nl

Subject
Assessment research project 2022-7553

www.ru.nl/etcgw

Dear Mr. van Halteren,

I hereby inform you that the Ethics Assessment Committee Humanities (EACH) of the Faculty of Arts and the Faculty of Philosophy, Theology and Religious Studies has evaluated the application of the research project *Mapping n-deletion in Dutch using tweets (application 2022-7553)* and has formulated the following advice on 21 February 2022:

This research project is approved for a period of five years (from 21 February 2022 to 21 February 2027).

Please note that any modification to the research project that might warrant review of the ethical approval must be submitted to the EACH.

Yours Sincerely,

ValidSigned door Marcel Becker
op 23-02-2022

dr. M.J. Becker
President of the Ethics Assessment Committee Humanities
Faculty of Arts and Faculty of Philosophy, Theology and Religious Studies