UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

WeST People and Knowledge Networks
Institute for Web Science and Technologies

"Did I Say Something Wrong?" A Word-Level Analysis of Wikipedia Articles for Deletion Discussions

Masterarbeit

zur Erlangung des Grades eines Master of Science (M.Sc.)
im Studiengang Web Science

vorgelegt von
Michael Ruster

Erstgutachter: Prof. Dr. Steffen Staab
Institute for Web Science and Technologies

Zweitgutachter: René Pickhardt
Institute for Web Science and Technologies

Koblenz, im Januar 2016

Erklärung

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe und die Arbeit von mir vorher nicht in einem anderen Prüfungsverfahren eingereicht wurde.

	Ja	Nein
Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	☒	☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	☒	☐
Der Text dieser Arbeit ist unter einer Creative Commons Lizenz (CC BY-SA 4.0) verfügbar.	☒	☐
Der Quellcode ist unter einer GNU General Public License (GPLv3) verfügbar.	☒	☐

.....

(Ort, Datum) (Unterschrift)

Abstract This thesis focuses on gaining linguistic insights into textual discussions on a word level. It was of special interest to distinguish messages that constructively contribute to a discussion from those that are detrimental to them. Thereby, we wanted to determine whether "I"- and "You"-messages are indicators for either of the two discussion styles. These messages are nowadays often used in guidelines for successful communication. Although their effects have been successfully evaluated multiple times, a large-scale analysis has never been conducted. Thus, we used Wikipedia Articles for Deletion (short: AfD) discussions together with the records of blocked users and developed a fully automated creation of an annotated data set. In this data set, messages were labelled either constructive or disruptive. We applied binary classifiers to the data to determine characteristic words for both discussion styles. Thereby, we also investigated whether function words like pronouns and conjunctions play an important role in distinguishing the two. We found that "You"-messages were a strong indicator for disruptive messages which matches their attributed effects on communication. However, we found "I"-messages to be indicative for disruptive messages as well which is contrary to their attributed effects. The importance of function words could neither be confirmed nor refuted. Other characteristic words for either communication style were not found. Yet, the results suggest that a different model might represent disruptive and constructive messages in textual discussions better.

Zusammenfassung Diese Arbeit beschäftigt sich damit, linguistische Erkenntnisse auf Wortebene über schriftlichen Diskussionen zu gewinnen. Die Unterscheidung zwischen Botschaften, welche sich förderlich auf Diskussionen auswirken und jene, welche diese unterbrechen, spielte dabei eine besondere Rolle. Hierbei lag ein Schwerpunkt darauf, zu ermitteln, ob Ich- und Du-Botschaften charakteristisch für die beiden Kommunikationsarten sind. Diese Botschaften sind über Jahre hinweg zu Empfehlungen für erfolgreiche Kommunikation avanciert. Ihre zugeschriebene Wirkung wurde zwar mehrfach bestätigt, jedoch geschah dies stets in kleineren Studien. Deshalb wurde in dieser Arbeit mithilfe der Löschdiskussionen der englischen Wikipedia und der Liste gesperrter Nutzer eine vollautomatische Erstellung eines annotierten Datensatzes entwickelt. Dabei wurden Diskussionsbotschaften entweder als förderlich oder schädlich für einen konstruktiven Diskussionsverlauf markiert. Dieser Datensatz wurde anschließend im Rahmen einer binären Klassifikation verwendet, um charakteristische Worte für die beiden Kommunikationsarten zu bestimmen. Es wurde zudem untersucht, ob anhand von Synsemantika (auch bekannt als Funktionswörter) wie Pronomen oder Konjunktionen eine Entscheidung über die Kommunikationsart einer Botschaft getroffen werden kann. Du-Botschaften wurden, übereinstimmend mit ihrer zugeschriebenen negativen Auswirkung auf Kommunikation, als schädlich in den durchgeführten Untersuchungen identifiziert. Entgegen der zugeschriebenen positiven Auswirkung von Ich-Botschaften, wurde bei diesen ebenfalls eine schädlich Wirkung festgestellt. Eine klare Aussage über die Relevanz von Synsemantika konnte anhand der Ergebnisse nicht getroffen werden. Weitere charakteristische Worte konnten nicht festgestellt werden. Die Ergebnisse deuten darauf hin, dass ein anderes Modell textliche Diskussionen potentiell besser abbilden könnte.

Contents

1. Motivation

Already in the 1960s, Thomas Gordon [19] coined the term *"I"-messages* while working as a psychologist with parents and their children. An "I"-message is described as an honest and direct message that expresses the sender's feelings. Its name is derived from them mostly having the first-person singular as subject. Furthermore, "I"-messages are characterised as being non-judgemental and thus e.g. suited for conflict resolution [12]. The concept became famous through Gordon's influential work introducing the *Parent Effectiveness Training* (short: *P.E.T.*) [21]. It describes an approach for parents communicating with their children for teaching them to solve their problems independently. Gordon [21] analogously calls messages with the second-person singular as subject *"You"-messages*. He attributes the opposite characteristics of "I"-messages to them, stating that they convey strong criticism. Gordon [20] later transferred these concepts onto leadership and reported many successful experiences from real-world applications. The general idea of "I"- and "You"-messages became widely used as recommendation for successful communication, even outside of leadership and teaching.

"I" and "you" are themselves *function words*, which is a set of common words that contribute little to the content of a sentence from a semantical perspective [8]. These include among others articles, prepositions, conjunctions, auxiliary verbs and other pronouns. However, Chung and Pennebaker [8] showed that function words, especially pronouns, can encode emotions and information that are not explicitly articulated in the text.

The role of function words in discussions has not yet been analysed. To our knowledge, the presumed effects of "I"- and "You"-messages have never been evaluated in a big scale on discussions as well. For these reasons, we wanted to analyse a great number of discussions for determining whether these concepts or certain words indicate "good" or "bad" communication. Mehrabian [32] found out that while communicating feelings and attitudes, the actual content merely makes 7% of the complete impression on the receiver of the message. The rest is shared by the tone of voice and gestures. To ensure that the analysed communication is only affected by the message contents, we considered purely textual discussions.

Due to the amount of available data, we used *Articles for Deletion discussions* (short: *AfD discussions*) for constructing a model. These are pages featuring textual discussions between Wikipedia users that determine whether an article should be deleted for reasons such as irrelevance or wrong information. Hereafter, messages in these discussions will be called *posts*. "Good" communication will be called *constructive* in this context. It can be described as positively affecting the collaborative process, e.g. by enriching the current discussion. Likewise, "bad" communication will be called *disruptive*. It is characterised by being highly detrimental to the collaboration process like posts that harass or attack other

users. In the context of AfD discussions, disruptive communication does not describe posts created with the intention to harm Wikipedia. Instead, disruptive messages should form an exception to authors otherwise trying to improve Wikipedia. To determine constructive and disruptive posts, we referred to a Wikipedia document containing records about users being (temporarily) blocked from editing due to misbehaviour. Thus, the task at hand was reduced to a binary classification problem and a gold standard based evaluation was conducted. Considering single posts, the dependent variable was whether such a post would lead to its author being blocked or not. As this yielded unsatisfying results, we additionally conducted tests in which the contents from a subset of the user's post history were considered.

We expected the results to broaden our understanding of which words are heavy indicators for messages in textual discussions being either constructive or disruptive. Of special interest were "I"- and "You"-messages and function words. Assuming that textual discussions in general and AfD discussions are comparable, the following research questions were derived:

RQ1: Do disruptive messages in textual discussions contain more "You"-messages than constructive ones? And likewise,

RQ2: Do constructive messages in textual discussions contain more "I"-messages than disruptive ones?

RQ3: Is solely considering function words sufficient for determining whether a message is constructive or rather detrimental to a textual discussion?

RQ4: Which other words are typical for constructive and which for disruptive messages in textual discussions?

Such knowledge could be used to encourage friendly collaboration and fight cyberbullying: When users use too many words that are common for disruptive messages, they could be asked whether they really intends to send the message. This could make them reconsider their wording which could otherwise be offending and detrimental to the ongoing discussion. In the context of the Google Science Fair 2014, a young researcher conducted a comparable study with encouraging results for adolescents to rethink sending offending messages [44]. Nevertheless, it is unclear, how effective such an approach would be for the mostly adult users of Wikipedia. Answers to these research questions could also be used to analyse AfD discussions on Wikipedia and alarm administrators when they should interfere to calm down a heated discussion.

Our results suggest that disruptive messages indeed contain more "You"-messages and thus support RQ1. However, the opposite, as questioned by RQ2, did not hold true for "I"-messages in our classifications. When only considering function words, the classifiers performed close to random guessing. This is possibly related to the full text classifiers' mediocre performance. Therefore, RQ3 cannot be answered although one of the tests returned surprisingly good results for function words classification. Inspecting typical words for disruptive and constructive messages did not return valuable results to answer

RQ4. Instead, these words are mostly specific to Wikipedia and AfD discussions. The results thus suggest that another model might be more suitable for analysing the language used in textual discussions than AfD discussions and the records of blocked users.

Chapter 2 introduces the concepts of "I"- and "You"-messages and function words in more detail, explaining their effects on and their role in communication. Related work, in which conflicts and disputes in textual discussions were analysed, is presented in Chapter 3. Moreover, approaches to detect vandalism on Wikipedia are shown. Chapter 4 presents the data we chose to build our model on, and the preprocessing steps that were needed to do so. Our fully automatic approach of building an annotated data set is explained in Chapter 5. In this chapter, we also introduce our classifiers and the metrics used in the later evaluation. Furthermore, characteristics of the newly created data set are studied. Chapter 6 describes the test setup with its goals and the chosen model validation. The results of our tests are given in Chapter 7. Chapter 8 illustrates assumptions and design decisions that we made and how they could have impacted the results. Finally, Chapter 9 concludes this thesis and the research questions are answered.

2. Psychological Concepts

This chapter focuses on the psychological backgrounds to this thesis. The sections further explain the concepts of "I"- and "You"-messages as well as function words. In both sections, studies are summarised which motivated us to inspect their impact within our model of textual discussions. Moreover, first details are given about how the AfD discussions will be tested for "I"- and "You"-messages and function words. Furthermore, it will be explained how their application to our model affects the research questions established in the first section.

2.1. "I"- and "You"-Messages

Gordon's conflict prevention and resolution recommendations do not solely focus on "I"- and "You"-messages. Instead, he highlights that his concept is based on two more skills besides these messages. The first skill is conflict resolution based on mutual agreement of both parties such that none of the parties feel to have "lost" the conflict [19, 20]. *Active listening* is described as a third important skill [19]. It advises one party to repeat the understood emotions and content of the other party's messages to ensure that both parties understand each other. Gordon recommends to refrain from using "You"-messages and instead use "I"-messages. The latter consist of three parts [20]:

1. A description of the intolerable behaviour of another party,
2. one's feelings towards this behaviour, and
3. the effect this behaviour has or will have on one's life.

The message should not be blameful. It should allow the listener to understand the speaker's feelings and the potential impact that the listener's actions may have on the speaker in general as well as on the speaker's well-being in particular.

A meta-evaluation by Müller et al. [34] confirmed many claimed positive effects of Gordon's training, especially for older children up to the age of twelve. Gordon [20] also highlighted successful experiences for applications in leadership. It indicates that Gordon's concept can also positively influence the communication and problem-solving between adults. Over time, the communication concept often got reduced to "I"- and "You"-messages and claimed to be an effective tool for general communication as well [45]. The messages were even further simplified so that the term "I"-messages was then no longer used for messages expressing behaviour, feelings and effects. Instead, they are nowadays mostly identified by containing the pronoun "I" while describing one's feelings towards someone else's behaviour. Yet still, studies e.g. by Kubany et al. [27] showed a

positive effect of the use of "I"- over "You"-messages. Further studies showed similar traits simply by inspecting the use of the pronouns "I" and "you". Weintraub [53] noted that angrier people tended to use "you" more frequently. A study on marital interactions saw "you" correlated with negative and "I" with positive effects on problem discussions [48]. Another study by Slatcher et al. [49] showed a positive effect of the more frequent use of "I" over "you" in couples' conversations on their relationship stability. However, the found correlations were less apparent.

All in all, many studies support the positive and negative effects of "I"- and "You"-messages and their simplified adaption. However, these studies are restricted to small group sizes or little text corpora. Hence, we want to determine whether we find evidence supporting Gordon's concept in AfD discussions. Creating an elaborate model to capture "I"- and "You"-messages would be out of this master thesis' scope due to its potential complexity. Instead, we reduce this task to counting of occurrences of the pronouns "I" and "you" similar to other research [e.g. 8, 48, 49]. The first two research questions can then be adapted to our model as follows: RQ1 poses the question whether disruptive posts contain the word "you" more frequently than constructive ones. Vice versa, RQ2 asks whether constructive posts contain the word "I" more frequently than disruptive ones.

2.2. Function Words

Function words span pronouns, articles, prepositions, conjunctions, and auxiliary words [8]. A common approach for analysing function words, as well as extracting implicitly encoded information from text in general, is counting words, which are believed to be connected e.g. with certain emotions. Pennebaker et al. [40] developed the frequently used program *Linguistic Inquiry and Word Count* (short: *LIWC*). It analyses text by counting words that it has internally mapped against emotions, content and function word categories. For example, words like "love" and "nice" can indicate positive emotions [41]. The leisure content category is associated with words such as "cook" and "movie" [41].

By conducting multiple studies using LIWC and investigating studies of other researchers, Chung and Pennebaker [8] found function words to encode different information. For example, Newman et al. [36] found that lying people were using function words measurably differently than truth-speaking people. Among other characteristics, liars were likelier to use fewer first-person singular pronouns and fewer exclusive words such as "but" or "except". A more frequent use of the pronoun "I" was also associated with higher blood pressure [8] as well as depression [8, 11]. Summarising three different studies, Chung and Pennebaker [8] found the use of pronouns to be an even better indicator for depression than the appearance of negative emotion words. Analogously, Chung and Pennebaker [8] attributed the frequent use of third person pronouns such as "she" and "they" positive effects on a person's well-being. By inspecting dialogues, the authors further found that people of lower status use the word "I" more frequently. Thus the relationship between two communicating parties may be subtly encoded. Most studies of Chung and Pennebaker [8] analysed English text of US-American participants. However,

when comparing results with English text that was translated from Japanese they found differences in the pronoun use and concluded that there are cultural differences in their use.

In sum, a person's use of function words and especially pronouns may unknowingly encode interesting information. In particular, "I" and "you" have been shown to correlate with the absence or existence of disputes. Due to the presumed effects of "I" and "you" and the broad range of information function words seem to encode, we hope to gain further insights and potentially detect new correlations. To our knowledge, no efforts have yet been made to determine whether connections between function words as a whole and constructive or disruptive communication in textual discussions exist. Therefore, we decided to not only do full text classification but classification exclusively respecting function words as well. The later classification may then give answers to RQ3, which asked if function words alone contain sufficient information to deduce whether a message is constructive or disruptive. As LIWC is a commercial software, we did not apply it to the AfD discussions. Nevertheless, results of studies conducted with LIWC highlight the importance of function words.

3. Related Work

In this chapter, research is presented that has already been carried out on conflicts on Wikipedia, Wikipedia discussions and textual discussions in general. Furthermore, information on research on *vandalism* detection is given, which is related to our work.

Vandalism is defined by a Wikipedia policy [W36] as the act of wilfully editing existing or creating new pages with the goal to harm Wikipedia. The policy highlights that, independent of the quality, any edits made with the intention to improve Wikipedia are not to be understood as vandalism. Therefore, articles that are obviously vandalism are not part of AfD discussions [W18]. Instead, they may be immediately deleted [W23]. This thesis aims to determine the impact of words in serious discussions that have the goal of reaching consensus. Hence, there is no interest in vandalism to us but only in posts made in good faith. We suspect the AfD discussions to attract few vandals because vandalising these pages causes relatively little harm to Wikipedia considering that not many people view them. Although limited, our experiences while working with these discussions support this assumption as we rarely saw any instances of vandalism. So, despite vandalism also being detrimental to rational and objective discussions, it differs from our goals to analyse good-faith discussions.

If not specified otherwise, the word *user* hereafter describes anyone who visits Wikipedia either for gathering information or for any form of contribution. A *contribution* is any active interaction made with Wikipedia such as writing articles, participating in discussions, reverting changes and the like. The word *glseditor* will be used to refer to the subset of users that contribute to Wikipedia, e.g. by writing articles or participating in discussions.

3.1. Conflicts and Disputes

This section presents research done on textual conflicts and discussions. The researchers' intentions varied from analyses to gain insights about the communication culture of users to the detection of disputes.

Our interest is to textually analyse posts to determine which ones no longer objectively address the relevance of an article and instead interrupt the discussion e.g. by containing personal attacks against other users. Yasseri et al. [56] on the other hand set out to detect and analyse conflicts during the collaborative editing process on the article itself. They have processed Wikipedia to detect collaboration conflicts called *editorial wars* (short: *Edit wars*). Edit wars are characterised by groups sharing different opinions trying to enforce their opinion in an article. Discussions related to the collaboration of an article happen on their respective talk pages. A talk page is a separate page dedicated solely

to the discussion of its associated article [W35]. The conflicts analysed by Yasseri et al. [56] take place on articles as well as their talk pages. The authors moreover set out to separate edit wars from pure vandalism and focus on serious discussions in good faith like we do as well. Edit wars are often accompanied by vandalism and reversions of the article to older revisions.

Deletion discussions end within a few weeks after a consensus was formed [W25]. Edit wars however often span longer or even an indefinite time. Yasseri et al. [56] distinguish three types of such conflicts. The first is when the war ends and a consensus is reached. A second type is identified by the authors as reoccurring phases of temporary consensus where a consensus seems to be reached multiple times but will always be broken up by another edit war phase. The last type is never-ending wars, which the authors identified to be typical for highly controversial topics such as the Liancourt Rocks [W13] that both Japan and Korea claim as their land. Yasseri et al. [56] did not find a correlation between the edit frequency of articles and conflicts.

To a reasonable extent, the length of the article's talk pages was an indicator for conflicts. However, the authors found it to be true for the English but not e.g. the Hungarian Wikipedia. Therefore, their approach to reliably detect conflicts in articles instead considers the number of editors E contributing to an article and a measure based on reversions [51]. Equation 3.1 shows the formula to calculate their controversy measurement M with the assumption that a high number of editors active in a discussion is an indicator for an edit war. N_i^d is the number of edits of the disputed article by the editor d of revision i. Revision i was reverted by the editor r of revision j for whom likewise N_j^r is the number of their edits. Although not captured by this formula, Sumi et al. [51] restrict the pairs (N_i^d, N_j^r) to those of *mutual reversions*. That is, the editor of revision i and j must both have reverted at least one revision of the other editor at some point. Thus, single direction reversions which are often a result of restoring a vandalised article to a prior state, are not weighted. Mutual reversions with at least one participant having vandalised the article contribute little weight due to considering only the minimum of each revision pair. The reason for this is that vandalism likely leads to an editor being blocked and thus they may no longer participate in a discussion for some timeframe [W21]. Conversely, conflicts between long-term editors are heavily weighted as the authors assume that such create more controversy.

$$M = E \cdot \sum_{(N_i^d, N_j^r) < \max} \min(N_i^d, N_j^r) \tag{3.1}$$

Ignoring the highest pair of mutual reversions prevents two heavily—perhaps even personally—fighting editors from skewing the result. Yasseri et al. [57] were able to successfully apply this measure to ten different language editions of Wikipedia—including such diverse languages as English, Arabic and Czech. They found the three most controversial categories to be politics, countries such as geographical locations or cities and religion.

While Yasseri et al. [56] detected and analysed the characteristics of conflicts in articles, Laniado et al. [29] set out to investigate editor interaction on article talk pages. For

every talk page, they built a tree with the article page as root and discussion comments as children. Replies to comments were treated as children to their original comment. Inspecting the height of the trees as well as the amount of nodes, Laniado et al. [29] determined extensively discussed categories. These categories are comparable to those found by Yasseri et al. [57] although Laniado et al. [29] only analysed the English edition of Wikipedia. Yet, as they chose different categories, Laniado et al. [29] additionally found philosophical and law related articles to attract lengthy discussions on their talk pages. In regard to the trees of such pages, this means that they contain many leaves with high depth. Furthermore, they found that editors, who reply to many others, mostly replied to inexperienced editors. Users who received many replies from others, frequently engaged in discussions with others. This analysis is restricted to a social networking level of interacting editors. In contrast, the goal of this thesis is the investigation on a content level instead of learning about the general Wikipedia communication culture.

Hassan et al. [23] developed a classifier to determine the attitude of Usenet users in threaded online discussions towards each other. They distinguish between positive and negative attitude. Among others, the first includes agreement and praise, whereas the negative attitude includes disagreement or insults. Their task therefore differs from ours in that we distinguish in disruptive posts and constructive ones. That is, insults often indicate disruptive posts. Yet, we regard disagreement as legitimate part of discussions and hence classify such posts as constructive as long as they do not contain personal attacks or similar. Hassan et al. [23] analyse data replies to other users extracted from various Usenet discussion groups. They constructed a graph from the words of each sentence in which they link words semantically related to each other or related by statistical co-occurrence. It is used to build Markov models, which are stochastic models. Those are used in a support vector machine (short: SVM) which achieves an F1 score of 80.2% and an accuracy of 80.3%. Both are evaluation metrics which are introduced in more detail in Section 5.2. The performance is notably better than our results.

Wang and Cardie [52] built a binary classifier for online dispute detection on Wikipedia talk pages. They performed a sentiment analysis on a sentence level and determined disputes according to the relation of positive to negative sentiments. Different to our approach, Wang and Cardie [52] predicted whether there is a dispute or not for whole discussions instead of single posts. The authors compiled a text corpus for testing by talk pages that are tagged with labels indicating disputes such as "`DISPUTED`". For non-disputed data, they referred to the absence of such tags on talk pages. Training took place on the *Authority and Alignment in Wikipedia Discussions* (short: *AAWD*) [3] corpus. The corpus consists of 365 discussions from talk pages that have been manually annotated by two or more annotators each. Relevant for the classifier developed by Wang and Cardie [52] are the *alignment moves* annotated by Bender et al. [3]. These are labels on a sentence level that categorise whether they express agreement or disagreement towards one or more other participants of the discussion. Besides the sentiment analysis, Bender et al. [3] also considered information about the discussion in total. Among others, their classifier additionally respected the length of replies, the number of editors active in the discussion and the topical category this discussion takes place in. Bender et al.

[3] note that arguments lead to longer answers. Similar to Sumi et al. [51], they argue that the participation of more editors in a discussion increases the likelihood of a dispute appearing. The authors furthermore made their classifier respect the category of the article being discussed as they argue that topics like politics or religion are likely to attract disputes. They used an SVM as classifier that respected all these features and returned an F1 score of 78.25% and an accuracy of 80.00%.

3.2. Vandalism Detection

Software tools that aid editors and especially administrators in detecting and removing vandalism play an important role in countering deliberate attempts of damaging Wikipedia [15]. Thus, developing algorithms that can accurately detect vandalism is an active field of research with manifold approaches of which a few are subsequently presented. Vandalism in posts is similar to disruptive posts in that both are detrimental to discussions. Yet, we assume good faith in what we call disruptive posts, whereas vandalism aims at wilfully harming Wikipedia. The presented vandalism detection solutions textually analyse individual posts' contents like in our approach. However, most also consider additional features such as metadata or structural information to improve their predictions.

Chin et al. [7] employ a bigram language model on article text to detect vandalism independent of any contributors. Their language model returns various measurements and statistics such as a post's number of words or perplexity, a commonly used measurement for evaluating language models. These are then classified by an SVM, logistic regression or decision trees. Instead of using a big data set, the authors tested this approach only on two of the most vandalised articles. They found decision trees to perform best. Moreover, Chin et al. [7] found little overlap in the vandalism detected by the SVM and logistic regression. Thus, they conclude that combined methods could further improve their vandalism detection approach.

Harpalani et al. [22] use trigram language models together with other features including the use of words from colloquial and vulgar language as well as objectiveness measurements. However, the authors saw the biggest prediction improvements when considering features based on a *probabilistic context free grammar* (short: *PCFG*). PCFGs are grammars whose rules are assigned a probability derived from the frequency of the rules being used in the training data [5]. Similar to a language model, Harpalani et al. [22] used a PCFG to detect a common writing style of vandals.

Adler et al. [2] on the other hand built a classifier respecting multiple features including a trust/reputation model of users. It calculates reputations per user and per country. The later is determined by the contributor's IP-address. Contribution metadata is another feature that the classifier by Adler et al. [2] respects. Among others, it considers the comment length and the time since the last edit. Extraordinarily short or long comments as well as frequently edited articles are an indication for vandalism [2, 54]. Another feature is the analysis of the contributed text for a high amount of uppercase letters, which vandals use for their posts to gain attention [33, 43]. The authors also considered

language features, namely the use of pronouns and words indicating biased contributions such as superlatives or bad style, e.g. colloquial language. Their motivation however is not based on the psychological effects of "I"- and "You"-messages but on their indication for non-objective contributions. This is rational considering that their focus was to detect vandalism on article pages where objective information should be compiled. As article pages provide contents that are of interest for the majority of people visiting Wikipedia, these are the main goal for vandals to gain attention and create great harm to Wikipedia. Therefore, current vandalism detection algorithms often focus on vandalism of article pages, whereas this thesis concentrates on discussions in which good-faith editors display misbehaviour. Nonetheless, the related topic of vandalism detection shows that there are many different features of Wikipedia contributions which successful classifiers may consider.

In conclusion, Wikipedia is actively used for discussion and interaction with other editors. The communication and interaction can remain peacefully constructive even over longer times. But there are also disputes of varying lengths of which some halt productive collaboration. In an effort to ensure good article quality and a minimum amount of article relevance, AfD discussions are intended as a tool for such peacefully constructive discussions. Nevertheless, some also transform into misbehaviour and are of particular interest to us. These incidents are per policy not vandalism as they are mostly based upon good-faith contributions that drift into personal attacks or harassment. Existing research on textual discussions and especially on Wikipedia discussions either focuses on the discussion as a whole or distinguishes between agreement and disagreement. In this thesis, however, we regard disagreement as a natural part of discussions and focus on linguistic characteristics of messages that are detrimental to collaborative processes due to e.g. verbally attacking other editors. Therefore, we solely consider text written by editors whereas other approaches often additionally or exclusively use structural approaches and other features. For example, networks were built from posts or edit and reply frequency were evaluated in previous research. To the best of our knowledge, this thesis marks the first efforts to evaluate textual discussions in a large scale for determining characteristics of language that indicate disruptive posts. Likewise, evaluations for the effect of "I"- and "You"-messages and function words in this setting were yet missing. The following chapters explain our approach from data preparation over test setups to results.

4. Data Extraction and Preprocessing

We processed the Wikipedia data dumps from the second of June 2015[1] which were the most recent at that time. They include all Wikipedia pages as well as the complete Wikipedia log, which documents actions including account creation, article deletion, blocking of users [W7]. Relevant for this thesis are the documented blocks. Hence, we filtered the log accordingly. Hereafter, the filtered log will be referred to as *block log* as it is done on Wikipedia as well [W1]. 326,538 AfD discussion pages and 3,332,551 registered occurrences of editors being blocked were found in the data dumps. The following Sections 4.1 and 4.4 go into detail about data properties of the AfD discussions and blocks respectively. Moreover, they describe our approaches and design decisions for preprocessing the data.

4.1. Articles for Deletion Discussions

AfDs are Wikipedia pages, which are used to discuss whether an article should be deleted. Many arguments revolve around whether an article fulfills the relevancy criteria [W29] of Wikipedia or not. That is, articles that are not deemed sufficiently relevant by the Wikipedia community will be deleted. Discussions often become heated due to disagreement between the two user groups called *inclusionists* and *deletionists* [W17]. Most of the time, an editor cannot be labelled as belonging to only one of the two groups but leaning towards one in certain topics. The inclusionists argue that most articles are worth keeping, i.e. should not be deleted, as long as they are relevant to a few people [W4]. This stance is motivated by the idea that contrary to a printed encyclopedia, the cost for one more article is negligible. Deletionists on the other hand argue that an article must be interesting to many people for fulfilling the relevancy standards of Wikipedia and can thus be kept [W4]. As a result, AfD discussions provide a great amount of human discussions featuring disagreement.

The procedure of an AfD discussion is as follows [W18]. If an article potentially fails to meet Wikipedia's quality standards, it may be nominated for deletion. However, if an article clearly validates Wikipedia's rules, e.g. contains copyright infringement or was unambiguously invented, it can be flagged for immediate deletion [W24]. An AfD nomination is normally done together with a short text describing why the article should be deleted according to the nominator. A dedicated AfD discussion page must then be created with the article's title as its title preceded by "`Wikipedia:Articles for deletion/`". Other editors may use it to discuss the quality of the article and to express their opinion about what should happen with the article in question. Besides keeping

[1] `https://dumps.wikimedia.org/enwiki/20150602/` – last accessed 13 July 2015, 15:20

or deleting the article, editors may also recommend actions such as merging it with another article. Usually, the AfD discussion pages are actively used for seven days to reach consensus on what action to take. Editors can only articulate recommendations, meaning that their statements are not votes and the discussions are not resolved by a majority decision. An administrator will consider all user-made recommendations and determine at their own discretion the most rational action that conforms to the Wikipedia policies. According to the guideline, the deciding administrator should not take part in the discussion and should not be involved in the discussed topic to prevent bias as much as possible. Nevertheless, it is not guaranteed that these guidelines are complied with by all participants.

4.2. Identifying and Isolating Posts

This section describes how individual posts by editors are obtained. The Wikipedia data dumps contain streamable, compressed archives of all pages together with all their revisions. There are no separate dumps that only contain AfD discussions. The pages and revisions are encoded as XML-files and AfD discussions are regular Wikipedia pages whose title starts with "`Wikipedia:Articles for deletion/`". Hence, we filtered the approximately 98 Gigabytes of archives according to the pages' title. We excluded summary pages which embedded old AfD discussions for documentation purposes. This was done to prevent duplicate data. The result was 29 Gigabytes of AfD discussions stored in uncompressed XML-files.

The data dumps do not include article modification histories. I.e. instead of computing and storing the differences between two revisions, every Wikipedia article revision consists of the complete text in this revision. Yet, for analysing editor posts on AfD discussions, it is crucial to separate them from the others. There are guidelines and recommendations for editors on how to communicate on discussion pages so that each editor's contribution can be clearly distinguished [W11, W31]. However, both guidelines and recommendations are not being enforced and are not always adhered to. Furthermore, Wikipedia pages are frequently being reverted to an earlier revision due to vandalism. Reverting editors inevitably introduce a lot of changes in the text which should not be attributed to them as they did not author the content. Likewise, if editors move text—e.g. a sentence—written by another editor to another place in the revision text, they should not be credited for this text. Thus, determining what text an editor introduced in some revision is a non-trivial task that must take all prior revisions into account.

Flöck and Rodchenko [14] addressed this problem by developing the algorithm *WikiWho* which associates words from a Wikipedia page revision with an editor that it assumes to be its author. The authors created a gold standard from Wikipedia articles which was then used for an evaluation in which their algorithm achieved a 95% precision. In so doing, WikiWho performed 10% better in this evaluation than the prior state-of-the-art algorithm by De Alfaro and Shavlovsky [10] while also executing faster [14]. Hence, we chose WikiWho for attributing authorship of words in revisions when extracting user posts. We only considered the contributions by registered users. This is motivated by

the fact that anonymous users are only distinguished by their IP address, which is not a unique identifier. A later presented approach merges posts by the same editor for the analysis and thus requires each post's author to be uniquely identifyable. Moreover, anonymous Wikipedia editors are likelier to vandalise pages than registered editors [25].

4.3. Removing Non-Linguistic and Automatically Generated Contents from Posts

After single posts have been extracted and those by anonymous users have been filtered, their contents have to be processed before they can be used in our later analysis. The purpose of this processing is to remove non-linguistic content like symbols as well as content that was not created by the post's author. Subsequently, we describe our removal of markup, templates and signatures.

As we are interested in determining which words negatively affect a constructive, objective collaboration, we want to ignore any forms of markup. Wikipedia allows the use of a subset of elements from the *HyperText Markup Language* in version 5 (short: *HTML5*) [W6] as well as *Cascading Style Sheets* (short: *CSS*) for formatting page contents [W5]. HTML element tags and attributes are removed, yet their contents are kept. CSS rules embedded in HTML attributes are therewith removed as well.

Wikipedia also supports a dedicated markup language called *wikitext* [W16]. Wikitext enables editors to format their text e.g. by using lists, making text bold or adding hyperlinks [W12]. It is converted to and rendered as HTML when a page containing this markup is viewed in a Web browser [W16]. Therefore, after determining an editor's contribution to a discussion, it must be processed to only contain words. The meaning of markup such as italics or bold is not standardised. Hence, for reasons of simplicity, we refrain from allowing the formatting to influence the weights of their affected words. However, it could be considered to e.g. attribute higher importance to a word in bold formatting and lesser to a struck through word in future work.

Naïvely removing all non-alphabetical symbols is insufficient for example in the context of marked up external hyperlinks. Hyperlinks are surrounded by square brackets like "`[http://ddg.gg]`". External links require one square bracket each, whereas Wikipedia internal links require two. Depending on the implementation, the removal of non-alphabetical symbols would result in one or more words being attributed to the editor as intentionally used in their style of communication. I.e. in this example it could falsely be assumed that an editor wrote the words "`http`", "`ddg`" and "`gg`". Especially URIs with long paths or query strings would add many words and thus, we ignore URIs marked up as links. Wikitext furthermore offers editors to provide a text for a hyperlink by using a vertical bar symbol "`|`". The text is shown instead of its URI when viewing the page e.g. "`[http://ddg.gg|only show this text]`". While we ignore the URI, as it is most likely beyond the control of the editors, any describing text can be freely chosen by them and will be taken into account in our setup.

After manually inspecting many AfD discussions, it became clear that links are often used to refer to internal Wikipedia pages like articles or policy pages. Linking to policy

pages is frequently done to draw another editor's attention to why their behaviour might be inappropriate for the ongoing discussion. These links are in the form of "`[[Wikipedia:No personal attacks]]`" and "`[[WP:NPA]]`". They can contain a hyperlink text similar to external links. We process internal Wikipedia links differently than external ones. The internal links beginning with "`Wikipedia:`" or "`WP:`" are retained and symbols including spaces removed. Thus, they will appear as one word to the classifiers, yet can be distinguished from regular communication contents. For example, "`[[WP:NPA]]`" would become "WPNPA". If there is a link text, it will be ignored as long as it is not different to the actual link.

Wikipedia templates [W9] also require special processing. They are used to easily insert commonly used text into pages such as page headers that describe the current state of an AfD discussion. As their contents have most likely been written by editors other than the editors who embed one or more of these in their posts, they should not be considered in our analysis. Templates can be included via transclusion [W10]. This is done by using two braces before and after the template name. For example, the "Like" template can be included by inserting the text "`{{Like}}`". The template's content will replace the template code "`{{Like}}`" when the page is rendered but the source code of the page will still show the template code. These templates are quite easy to detect and are removed during our content processing steps.

However, templates can also be substituted [W8]. The referred template's content is then inserted in its most recent revision into the source code of the page. For example, the "Like" template is substituted by adding "`{{subst:Like}}`" to the page. Most AfD-specific templates are to be substituted [cf. W18, W33]. The task to detect substituted templates is thus non-trivial: All revisions of every template existing prior to the creation of the page in question would have to be matched against the full text of all AfD discussion revisions. That is, a template detection would have to be applied before the WikiWho algorithm, because WikiWho does not guarantee to reconstruct revisions correctly. The algorithm moreover discards any spaces. We refrained from implementing an algorithm which would detect any substituted templates due to the associated computational overhead. Beyond that, most templates are rarely used or even substituted in AfD discussions in our experience. This is because templates are mainly embedded in other spaces than in AfD discussions such as the user or article space; cf. the list of Wikipedia templates [W3]. Nonetheless, there are a few templates built to be used in AfD discussions [W2]. They are mainly used to open such a discussion or state that it has been finished together with its outcome. As they are a recurring part of AfD discussions, we have implemented a rudimentary detection and removal of these templates. Due to the complexity of this task, attempts to automatically build detection mechanisms from the revisions of all AfD templates without heavily impacting the performance failed. Instead, we identified the most commonly used templates in AfD discussions posts by grouping identical posts and counting them. Thereafter, we built regular expressions to match and remove frequently used AfD templates in posts.

As per guideline [W31], many editors append a signature to their posts. There is also a bot called SineBot [W14] dedicated to adding signatures for editors who forgot

to add or purposely left out their signature. Posts by this bot alongside with other posts by registered editors whose user names ends in "Bot" are ignored because we are interested only in human communication. This action follows the naming scheme described in the Wikipedia user name policy [W34]. Signatures should be removed for two reasons. First, they can be customised by registered users and thereby introduce new words to their posts, which could influence the results. Second, customised signatures as well as signatures in general can be used to identify authors. However, classifiers should not learn to detect editors and decide upon their behaviour but make predictions only on the actual discussion contents. Consequently, we match and remove all regular signatures as well as many customised ones. Due to the vast possibilities of customising signatures using wikitext, HTML and CSS, we settled for a removal of many but not all customised signatures. This design decision and its potential effects are further discussed in chapter 8.

Performing the removal of automatically generated and non-linguistic contents after applying the WikiWho algorithm was performance motivated: Without identifying single contributions first, the removal would have had to be done on the most recent and all previous contributions for each revision.

4.4. Preprocessing of blocks

Besides extracting and preprocessing the posts of AfD discussions, similar has to be done for the information about blocked users retrieved from the block log. At first, all blocks issued on anonymous users were removed as these are distinguished only by their IP address which is not unique. This is done analogously to our post removal by anonymous editors. Without this precaution, multiple people could issue edits from the same IP address, with some being constructive and others disruptive. As a result of the filtering, the number of blocks was reduced from $3,332,551$ to $821,074$.

Per block, the block log contains a timestamp, the blocked user name, the user name of the administrator issuing the block together with the administrator's internal user ID, and a comment. As we ignore anonymous editors, the user names are unique. Hence, it is not a problem that an internal ID for blocked users is missing. Each comment should legitimise its associated block. We use these comments to determine whether a block is relevant for our analysis. The comment can be freely chosen by the blocking administrator or even left blank. Therefore, some comments are more helpful than others when trying to determine which blocks are of interest. For example, comments such as "personal attacks" or "harassment" are the result of disruptive contributions in which an editor must have verbally attacked another user. Conversely, "Willy" as a reason cannot be understood without a context. Here, the comment refers to the user "Willy on Wheels" who is claimed to have created more than a thousand accounts for the purpose of vandalising Wikipedia pages [W15]. Judging from the block log alone, these claims seem realistic.

There are two options to determine which blocks should be considered based on their comments. One is a whitelisting approach, where only blocks are extracted that match

a set of words which are indicative for preceded disruptive contributions. The other option is to blacklist words that indicate prior misbehaviour unrelated to an editor's style of communication. For whitelisting, we considered the words "personal attacks", "harassment", "vandalism", "hating" and "legal threats". We included "vandalism" as vandalism has disturbing effects on discussions and is thus more similar to disruptive than to constructive posts. As the comments are freely chosen texts by administrators, they may incorrectly claim a disruptive post to be vandalism because on first sight, both might look similar and may result in a block. Wikipedia shortcuts [W30] were also considered. These are internal links that in this case link to the respective policies. For example the policy explaining that personal attacks should be avoided and may lead to blocks [W28] is abbreviated using the shortcuts "WP:NPA" and "WP:PERSONAL". Even with additionally respecting word stems such as "vandal" or "harass", the whitelist approach resulted in merely 16.548 blocks issued on registered Wikipedia editors. The number does not take into account that some of the editors will not have been active in an AfD discussion at all. Therefore, even fewer data on blocked editors would be available. Furthermore, it disregards more freely worded and context-sensitive comments such as "`being a dick after being blocked for the same thing`", where "the same thing" refers to an earlier block issued due to "attacks, incivility" by the same administrator on the same editor.

For the blacklisting approach, only those blocks were removed which were unlikely to be issued because of a disruptive communication style by humans. Blocks whose legitimation comment contained the word "bot" were removed as our interest lies in human communication and not e.g. in bots malfunctioning. Naturally, blocks that were issued only for testing purposes were also ignored. Sometimes editors want to force themselves to take a break off contributing to Wikipedia and hence request being blocked [W20]. As the motivation behind such a block is unlikely to be related to disruptive editing, we ignore it as well. Blocks that were issued due to the contents being spam or advertisements were also ignored. We do not aim to build spam detection but instead we are interested in the used language that indicates misbehaviour such as personal attacks or legal threats. All blocks of editors that concerned copyright infringement have been removed. This was done, because these claims cannot be automatically detected without matching the contributions against a great amount of external data sources. Moreover, the detection of copyright infringement is not the goal of this thesis. Finally, we discarded the blocking of editors associated with not citing any or solely untrustworthy sources. This misbehaviour must have happened on actual article pages which we do not consider and thus a certain style of communication on AfD pages cannot be derived from it. The blacklisting approach returned $800,737$ blocks and also yielded better results for all classifiers.

We determined $746,349$ registered editors that have been blocked at least once according to the block log with respect to our previously described criteria. On the other hand, $25,743,749$ registered users [W38] have never been blocked. In other words, merely 2.82% of all registered Wikipedia users have been blocked at least once. But of the $25,743,749$ users, there are also users who registered an account but have never been active. However, considering all $111,759$ editors who have been active in at least one

AfD discussion, 16,314 or 14.6% of all AfD participants have been blocked at some time. Therefore, the AfD discussions offer a large pool of human textual discussions that partly created enough arousal for an administrator to intervene and block an editor.

5. Building an Annotated Data Set

For learning and testing classifiers, we need an annotated data set. That is, we need posts that are labelled as disruptive or constructive. However, none of the entries in the block log is associated with a user contribution. Thus, it is unclear, what contribution led to a block being issued. Moreover, such a data set did not yet exist for AfD discussions. The existing annotated data set AAWD [3] was built from general Wikipedia talk pages. It contains merely 3000 additions of text and was annotated using data from 2008. However, to prevent overfitting we were interested in a bigger data set. It is also unclear, whether the language of discussions has changed within the years and thus, whether annotations from 2008 still represent current discussion styles. At least when considering the oldest AfD discussions, we found that our classifiers perform notably different as later presented in Section 7.3. Moreover, regarding the grading scale by Landis and Koch [28], the annotators only showed "moderate agreement" with a Cohen's kappa coefficient of 0.50. In addition, only agreement and disagreement were labelled. A subset of what Bender et al. [3] identified as disagreement, matches our definition of disruptive posts. Therefore, we decided against using AAWD.

Manually annotating posts was not an option, because it would have only been feasible with a small subset of data. Yet, we were interested in gathering information from the analysis of large-scale textual discussions. We assumed that disruptive behaviour must have preceded a block. In this chapter, we describe how we generated an annotated data set of constructive and disruptive posts with this assumption in mind.

Assuming that either all or only the last n posts by editors before they were blocked led to said block, may result in false predictions. For example, there could exist a scenario in which formerly constructive editors took a longer break from editing and then were blocked due to their new contributions. As they were not blocked for their old posts, it is unlikely that these had been disruptive. Yet, the old posts could then be classified as such in our data set. Therefore, we chose to assume that the posts made by editors shortly before they blocked were the cause of it. This raises the question: How much time should be considered before and up to a block, in which we assume posts by the blocked authors to be disruptive? Hereafter, we call this timespan a *timeframe*.

For every post, we calculated the time between the creation of this post and the next time its author was blocked. Posts whose authors were never blocked afterwards, were considered to have been constructive regardless of any applied timeframe. This data was then used to run tests using different timeframes and compare the performance of our full text classifiers. The classifiers are described in more detail in Section 5.1. In the then ensuing section, the metrics we chose for evaluating classifier performance throughout this thesis are presented. Successively, the test setup and results for determining the best

timeframe are shown. Finally, some analysis is done on the newly generated annotated data set without the use of classifiers but by inspecting the term occurence frequencies.

5.1. Classifiers

For choosing a timeframe and for our later analyses, we decided to use a support vector machine (short: SVM), a naïve Bayes (short: NB) classifier and a language model (short: LM) classifier. Regarding the NB classifier and support vector machine, two commonly used classifiers have been selected. SVMs are known to be well-suited for text classification and to perform better than NB classifiers [24, 55] in this context. Therefore, we expected to see similar performance differences for our corpus. The language model classifier was chosen because it fits our analysis well in which we want to better understand language used in discussions. Both, the LM and the NB classifier, are probabilistic models. They are the same when considering their simplest implementations [cf. e.g. 31, 39]. Therefore, we expected them to perform at least comparably to each other. Yet, our LM classifier does not consider words independently like an NB classifier but instead uses a small history of words. Due to inspecting words in a local context, we suspected that an LM classifier may also outperform an NB classifier. However, the results, as later presented in chapter 7, show that the NB classifier performs better than the LM classifier. The SVM on the other hand did indeed outperform the other classifiers.

After preprocessing, the posts containing English words, markup, numbers, punctuation marks and other symbols had been reduced to words separated by one space symbol each. We declare the set of terms as the vocabulary $V \coloneqq \{t_1, t_2, \ldots, t_n\}$. In this sense, posts are elements of the set of sequences $S \coloneqq V^*$ constructable from all terms. A particular post will be denoted as sequence $s = (w_1 w_2 \ldots w_m)$ with a word $w_i \in V$. Thereby, it is possible for $w_i = w_j$ with $i \neq j$, i.e. a post can contain multiple occurrences of the same term. The set of sequences on which the classifiers are learnt will be referred to as S_L and the one on which the classifiers are tested as S_T. Thus, it holds that $S_L \subset S$ and $S_T \subset S$ with the possibility of there being an overlap between S_L and S_T. Such an overlap is likely for typical AfD discussion posts such as "`delete as per nom`", which expresses that the author agrees with the proposed deletion and the reason given by the creator of this AfD discussion. The vocabularies containing the unique terms from the sequences on which the classifiers are learnt and tested are denoted as V_L and V_T respectively.

The SVM and the NB classifier were both given each post as a *term frequency-inverse document frequency* (short: *tf-idf*) feature vector. A high tf-idf value indicates an extraordinarily high presence of a term within a post (term frequency) compared to its appearance in the complete corpus (document frequency). When a term appears in many posts, its document frequency is high. Using the inverse ensures that the tf-idf value shrinks consequently to penalise common terms like conjunctions that do not represent a post's content well. This allows retaining stop words in posts, which are essential for the function word analysis, while reducing their otherwise heavy impact. In small tests, we found little difference between the performance of either using a tf-idf or a tf feature vector. In the context of RQ4, we decided to use a tf-idf feature vector for finding terms

other than function words that are characteristic for disruptive or constructive messages. This did not interfere with determining the effects of function words, as required by RQ3, because separate classifications that solely considered function words were done as well.

Given the great amount of posts to analyse, the feature vectors for the SVM and NB classifiers have a high dimensionality while being sparse. Thus, we applied stemming e.g. to reduce conjugated verbs to their word stem. In many different classification tasks, very short words like one letter words or stop words are being pruned to reduce the vector size. We refrain from doing so to ensure that the effect of function words, which include stop words and are often very short such as "I" or "a", can be studied. RapidMiner's snowball stemmer was used for the task of stemming as its result were slightly better than the Porter and equal to the Lovins stemmer which are also included in RapidMiner. The following sections present the general concepts of our chosen classifiers individually together with some design choices we made. We used RapidMiner[1] Studio 5.3, a software frequently used for machine learning and data mining tasks, for our SVM and NB classifications. The LM classification process was developed using the Generalized Language Model Toolkit[2].

5.1.1. Support Vector Machine

Support vector machines [9] are supervised learning models for solving binary classification tasks. They treat elements of a class—in our case each represented by a tf-idf feature vector—as points in a high-dimensional space. Using training data, an SVM tries to calculate the optimal hyperplane h for separating the points of both classes as shown in Figure 5.1. A hyperplane is optimal when it maximises the margin between the elements of each class closest to the hyperplane. Only the three circled elements in the left plot of Figure 5.1 determine the margins and as a consequence thereof the hyperplane. Thus, they are called *support vectors*.

In many cases, the elements of the two classes are not linearly separable without errors. Then, a soft margin hyperplane can be used that tolerates a minimal amount of errors. This is shown in the right plot of Figure 5.1. In cases where the data is not linearly separable as depicted in the left plot of Figure 5.2, Cortes and Vapnik [9] propose the transformation of the vectors into higher features spaces. The right plot of Figure 5.2 illustrates this by using a mapping function ϕ. Instead of transforming all points into a higher feature space, kernel functions are used. They can calculate a dot product of two points in a higher dimensional space as needed for determining the hyperplane as well as for the actual nonlinear classification. The interested reader may refer to Cortes and Vapnik [9]. Our classifier uses a dot kernel $k(u, v) = u \cdot v$, say the inner product of u and v. Due to limited time for this thesis, other kernels were not tested.

[1] `https://rapidminer.com/` – last accessed 22 January 2016, 15:30

[2] `https://github.com/renepickhardt/generalized-language-modeling-toolkit/tree/cebbff8` – last accessed 22 January 2016, 15:30

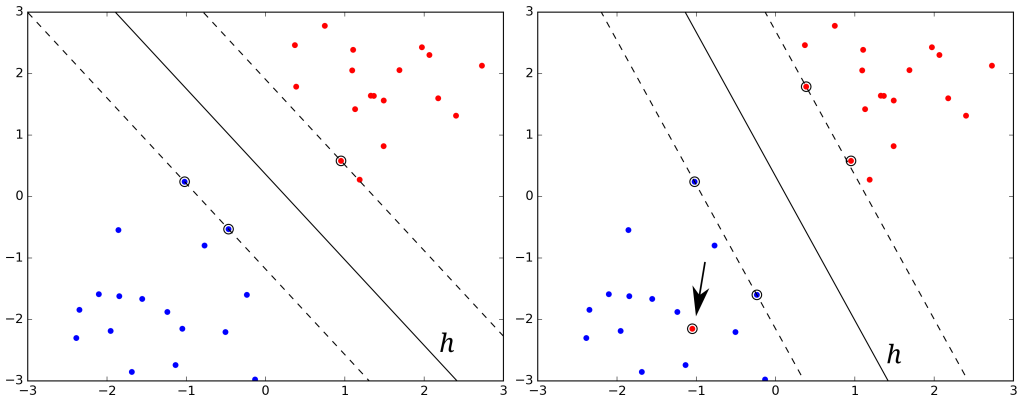

Figure 5.1.: Two elements of different classes are separated by an optimal hyperplane h as the margins (dashed lines) are maximised. Circled elements are support vectors. In the right plot, the use of a soft margin allows tolerating an error, indicated by the black arrow.

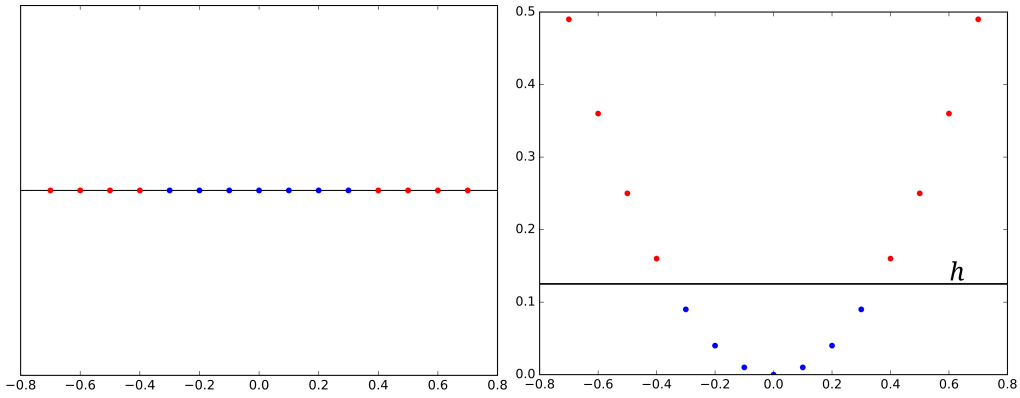

Figure 5.2.: The left plot shows a dataset of two classes which are not linearly separable in a one-dimensional space. In the right plot, the data has been transformed into a two-dimensional space by using a mapping function $\phi\colon \mathbb{R} \to \mathbb{R}^2,\ x \mapsto (x, x^2)$. The classes can thus be separated by a hyperplane h.

5.1.2. Naïve Bayes Classifier

A naïve Bayes classifier is a probabilistic classifier. We use a multinomial NB classifier in this thesis. Such a classifier is a unigram model, meaning that it determines probabilities for words individually instead of considering their usage context [31]. Equation 5.1 shows the general concept of a multinomial NB classifier [cf. e.g. 31] using tf-idf weights [46].

$$\gamma_{\mathrm{NB}}\left(s \in S_T\right) = \underset{k \in \{0,1\}}{\arg\max} \left(P\left(C_k\right) \cdot \prod_{w \in s} P\left(w \mid C_k\right)^{\text{tf-idf}(w)} \right) \tag{5.1}$$

The classifier is applied to a specific post s from the testing data with $C \coloneqq \{0, 1\}$ being the prediction classes where 1 indicates a post that led to a block and 0 one that did not. Essentially, one classifier is learnt per class and their predictions for s are compared against each other. The argmax operator returns the class of which the words in s had the highest probability to belong to. $P\left(C_k\right)$ is the general probability for any sequence to belong to class C_k. Using balanced data, the probability is $p = 0.5$.

A zero probability would result in the whole product becoming zero. Due to the sparsity of the feature vector, zero probabilities are likely. Laplace smoothing solves this problem by adding 1 to the denominator and the number of unique words in the training data $|V_L|$ to the nominator [4]:

$$\widehat{P}_{\mathrm{Laplace}}\left(w \mid C_k\right)^{tf(w,s)} = \frac{\mathrm{tf}\left(w, S_{L_{C_k}}\right) + 1}{\sum_{v \in V_L} \left(\mathrm{tf}\left(v, S_{L_{C_k}}\right) + 1\right)} \tag{5.2}$$

For illustration purposes, Equation 5.2 uses term frequency instead of tf-idf for explaining an estimation of $P\left(w \mid C_k\right)$ using Laplace smoothing. $S_{L_{C_k}}$ is the set of learnt sequences of class C_k. The term frequency $\mathrm{tf}(x, S_{L_{C_k}})$ is the summed up occurrence frequency of the term x in all sequences of $S_{L_{C_k}}$. In this estimation, the term frequency of w is set into relation to that of the learnt terms $v \in V_L$ while adding 1 to prevent a result of zero [26]. With big enough training data, the addition's effect on the estimated probabilities becomes negligible. Another option would be Lidstone smoothing which is a generalisation of Laplace smoothing using any value instead of just 1. However, using Lidstone smoothing for the NB classifier was not an already available option in RapidMiner.

5.1.3. Language Model Classifier

The third classifier uses the 4-gram language model with modified Kneser-Ney smoothing introduced as generalised language model by Pickhardt et al. [42]. An n-gram language model estimates a probability for a sequence. It does so by analysing each word's probability given the local history of the last $n - 1$ successive words prior to the current word. It is therefore a probabilistic model. The decision to only consider n words of a sequence follows a Markov assumption so that full sequences' probabilities are only approximated and thus computational efforts are reduced. Simultaneously, it is a necessity

as it is impossible to have learned all potential sequences. A trivial language model will therefore calculate a sequence's probability by calculating the conditional probabilities of the words $w_i \in s$ as shown in Equation 5.3.

$$P(s \in S_T) = \prod_{i=1}^{|s|} P(w_i \mid w_{i-n+1} \dots w_{i-1}) \tag{5.3}$$

Such a trivial implementation however would result in zero probabilities for unseen terms. Likewise, the confidence of estimations on rare terms would be low. Both cases are likely to occur, considering that human languages follow Zipf's law and thus most words in a text corpus appear only a few times [35]. To overcome the problems introduced by data sparsity, smoothing techniques are applied. The language model applied to the AfD discussions uses a modified Kneser-Ney smoothing to solve this task. Essentially, the smoothing interpolates higher order with lower order models; using relationally growing discounts for n-grams that occur more frequently. The interested reader may refer to Chen and Goodman [6].

Pickhardt et al. [42] embodied another concept called *skip n-grams*. Besides the interpolation with lower order models as is done in modified Kneser-Ney smoothing, wildcards are applied to n-grams. A 3-gram model would thus determine $P(w_3 \mid w_1 w_2)$ by not only interpolating with $P(w_3 \mid w_2)$ and $P(w_3)$ but also with $P(w_3 \mid w_{1\_})$ where the "_" symbol indicates a wildcard. This reduces data sparsity problems for higher order models and hence allows a language model to learn relations between words which are not successive [18]. For example, if a 3-gram model would be learnt on the sentences "I love you" and "I like you", skip n-grams would enable it detecting a relation between "I" and "`you`" separated by a single word. The model thus learns a context and would calculate a higher probability for the previously unseen sentence "I envy you" than a model that does not use skip n-grams.

Similar to the approach of the multinomial NB classifier, we train two separate language models. One is learnt on posts that led to a block and another on posts that did not. The class prediction of a statement s is then made by determining the better performing model for it. To evaluate the performance, we use the perplexity e^H with H as given in Equation 5.4 where $\widehat{P}_{\text{GLM}}$ is the generalised language model by Pickhardt et al. [42]. Perplexity [16] is a common metric for evaluating language models with lower values indicating a better model.

$$H(s \in S_T) = \frac{-\sum_{i=1}^{|s|} \log\left(\widehat{P}_{\text{GLM}}(s)\right)}{|s|} \tag{5.4}$$

5.2. Evaluation Metrics

Accuracy, precision, recall, area under the curve (short: AUC) and F score were chosen as evaluation metrics. They are commonly used for binary and text classification evaluation [50]. This section introduces these metrics and explains how they will be utilised in our classification evaluations.

Accuracy is the ratio of correct predictions compared to the number of all predictions. It is the only of our chosen metrics that incorporates the correct prediction of both disruptive and constructive posts. The others can be calculated for positive predictions—posts that led to a block—as well as negative predictions which are posts that did not lead to a block. This thesis' focus is the detection of disruptive posts for determining the effects that function words and "I"- and "You"-messages have on this task as well as other terms. Thus, the positive predictions are of special interest. Positive precision is the probability that a post, which has been predicted to be disruptive, truly was disruptive. The negative equivalent is often also called false omission rate and is calculated analogously.

Positive recall is the probability that given a disruptive post, it is predicted as such. Negative recall is also called fall-out and is calculated as the probability that given a constructive post, it is predicted as such.

In accordance with our goals, a high positive precision is more important than a high positive recall for discovering language characteristics. That is, a classifier might predict fewer disruptive posts but these predictions would then be correct more often. Therefore, characteristic terms of disruptive posts can rather be retrieved from a classifier with high prediction.

A high precision or recall can easily be achieved by a bad classifier: If a classifier only predicts a single disruptive post but does so correctly, it will achieve a positive precision of 100%. Likewise, if a classifier predicts every post to be disruptive, it will achieve a positive precision of 100%. However, these two perfect values negatively influence each other and cannot coexist when a classifier uses this naïve approach. To capture this, the F score can be used as it incorporates both values. We use the commonly utilised F1 score which is the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{5.5}$$

Consequently, there is a positive and a negative F1 score.

The AUC is the area under the receiver operating characteristics (short: ROC) graph. This graph's x-axis contains the false positives rate whereas the y-axis contains the true positives rate. All points are sorted by prediction confidence from highest to lowest. Whereas the other metrics are highly affected by the class distribution, ROC curves are insensitive to heavily skewed distributions. Naturally, when balancing the data, this benefit becomes irrelevant. The AUC describes a classifier's ability to correctly predict disruptive posts while ignoring any constructive posts. Theoretically, one could also plot false negatives on the x-axis and true negatives on the y-axis. However, this is uncommon and not in our interest. Therefore, we do not distinguish between a positive and a negative AUC. The interested reader may refer to Fawcett [13] for an extensive description of both ROC and AUC.

Determining the performance of the classifiers with different timeframes is done on balanced data. Likewise, all later tests but one are executed on balanced data as well. Thus, accuracy will be used as metrics for the overall performance of classifiers. As for judging the performance of classifying disruptive posts, the positive F1 score is of special

interest. Similarly, the AUC will give information about the performance of disruptive post classification in terms of prediction confidence.

5.3. Deciding on a Timeframe

For constructing an annotated data set, a timeframe must be chosen. Multiple tests were run using our classifiers on data with different timeframes. The arithmetic mean of their performances was evaluated to ascertain the best timeframe length. As the classifiers run for multiple hours and as there are many testable timeframes, our testing was limited to a few timeframes. In this section, we explain which timeframes were chosen to be tested and which of these led to the best overall performance.

To not overfit the data and thus to allow detecting general communication patterns, we were looking for timeframes in which at least $10,000$ blocks were issued on editors who contributed to an AfD discussion at least once. Consequently, the lowest timeframe was set to 13 hours in which there were $10,067$ such blocks. We inspected the last post an editor had made in an AfD discussion before they were blocked with respect to the post's temporal distance to said block. Figure 5.3 shows how a growing timeframe increases the number of these last posts which would be considered disruptive. The number of posts quickly rises in the beginning with its growth speed decreasing. This suggests that editors who disruptively act in AfD discussions are blocked soon after their last disruptive posts. According to the corresponding Wikipedia policy [W21], blocks should only be issued to ensure a productive and disruption-free environment. As such, they should not be used for punishing misbehaviour of users and thereby should not be issued on old posts which no longer affect active discussions. Thus, the longer the considered time between the last post by an editor in an AfD discussion and their block, the likelier it is for the block to be related to a contribution elsewhere on Wikipedia. AfD discussions should typically be closed after seven days [W18]. Therefore, 6 days was chosen as our longest timeframe for testing, so that an AfD discussion would still be active and benefit from an editor being blocked. The other timeframes were decided to be 1, 1.5, 2, 2.5, 3, 4 and 5 days long following the assumption that longer timeframes will eventually lead to worse data quality.

Within a timeframe of 13 hours, only 0.70% of all posts are disruptive. A classifier that would always predict a contribution to be constructive would yield a good overall performance. Thus, the data was balanced to prevent this effect. As a consequence thereof, the amount of data a classifier has to process is reduced immensely which makes the process faster. Otherwise, the unbalanced data would have to be sampled to a feasible subset. Due to the then small amount of disruptive posts, these would likely have been overfitted.

To make the results between different timeframes more comparable, the classifiers were always learnt on the same sample size. With longer timeframes, the number of assumed disruptive posts increases at the cost of the amount of constructive posts. Therefore, all available disruptive posts for the timeframe of 13 hours were used and for later timeframes, the same amount was randomly sampled.

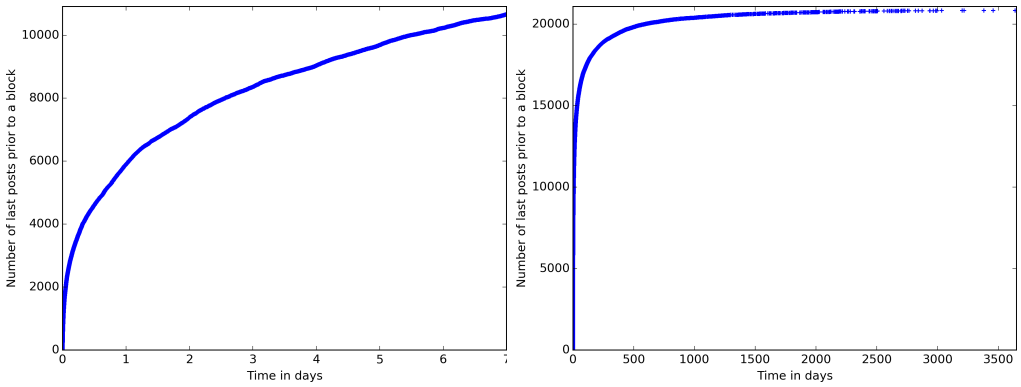

Figure 5.3.: Number of last posts in an AfD discussion prior to their authors being blocked within a given timeframe in days. The left graph shows the first seven days of the right graph in more detail.

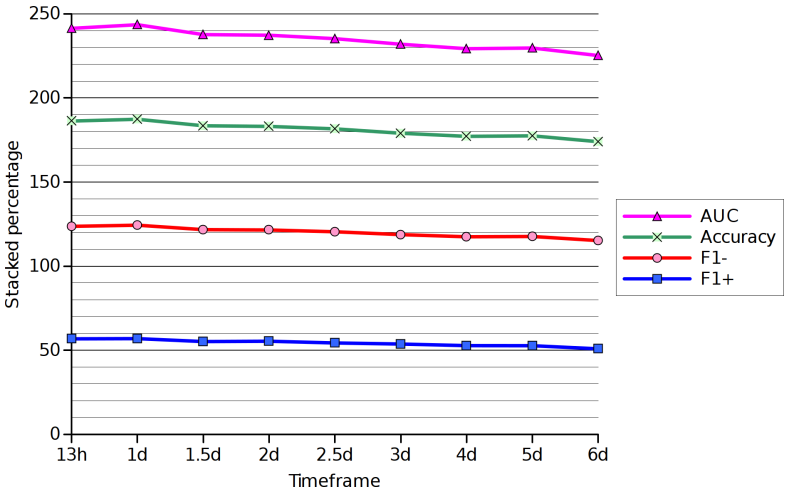

Figure 5.4.: This stacked line plot shows the average performance of the classifiers with the various timeframes. The values are the arithmetic mean of the results from the SVM, NB classifier and LM classifier. All are given as percentages.

As can be seen in Figure 5.4, the timeframe of 1 day returned the best results overall. The figure shows how the values for all metrics peak at the 1 day timeframe and worsen the longer the timeframe becomes. All specific values used to generate the plot in Figure 5.4 can be found in Table A.1 together with precision and recall values. The positive precision value is the only one which was not the highest using a 1 day timeframe. It was 49.81% for 13 hours and 49.29% for 1 day. Nevertheless, this is negligible considering that the 1 day timeframe resulted in the greatest positive F1 score and accuracy. Thus, successive tests were performed using 1 day as timeframe length.

Due to the missing link between contributions and blocks, we made the assumption that posts shortly before a block were disruptive. However, there are scenarios where this assumption is false. This problem will be discussed in chapter 8.

5.4. Data Analysis

Besides through classification, the data was also manually inspected. For this, all AfD discussions have been separated into single posts, which have been preprocessed, e.g. by removing wikitext markup. All posts have been labelled as either being disruptive or constructive according to the 1 day timeframe.

First, the average length of disruptive and constructive posts was compared. The length was defined as the number of words a post contained. We considered all posts labelled disruptive within the 1 day timeframe and randomly sampled the same number of constructive posts. Here, we only discuss concrete values of a single subset of all available posts because of performance reasons. Analysing and then plotting the length of $3,467,402$ posts was not an option. Nonetheless, we repeated this test with different samples and found similar results. In general, for every data set we found the average post length always to be significantly longer for disruptive posts. The average disruptive post is 23.80% longer than a constructive one. That is, the average disruptive post contains 41 words, whereas the average constructive post only contains 31. However, with the medians being 17 and 15 respectively, the data is heavily skewed, i.e. it contains multiple outlier posts with many words. The box plot in Figure 5.5 illustrates this.

In general, shorter posts seem to be less disruptive than longer ones. An explanation could be that many short posts solely state to keep, merge or delete the article. They often lack an explanation why the article should be kept, merged or deleted. Such a post is an unfounded expression of one's opinion about the quality and relevance of an article. Naturally, this alone is no reason for being blocked. Personal attacks and similar disruptive comments can also be made using few words but the data suggests that such posts are likelier to be verbose.

Table 5.1 shows how often certain terms appeared in the disruptive or constructive posts in relation to all other words of the posts. All $3,467,402$ posts have been considered for these values. Common swear words like "fucking", "fuck" and "shit" are hardly used in disruptive posts. With only 6.39‰ of disruptive posts containing any of the three swear words, the few posts that do, contain the terms multiple times. However, when they are used, they are quite expressive. For example, the term "shit" is 9.43 times

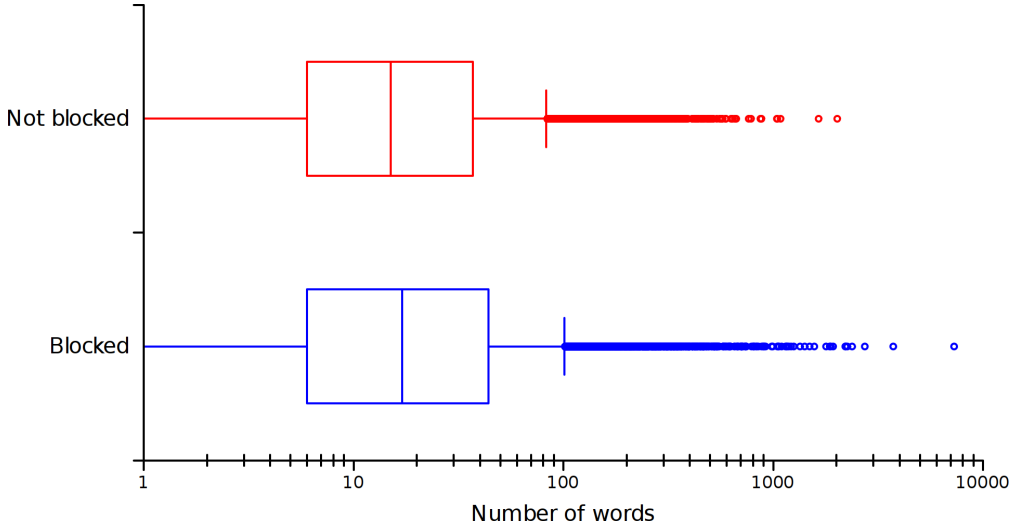

Figure 5.5.: This log-scaled box plot shows the differences in lengths of posts which were assumed to have led to a block and those which were not. It illustrates that both data sets are skewed and have outliers consisting of many words. In spite of that, disruptive posts contain more words on average.

likelier to appear in a disruptive than in a constructive post. In sum, with these swear words alone, a small recall but a high precision could potentially be achieved. Hence, a collection of swear words would be unlikely to suffice for identifying many disruptive posts.

The use of "I"- and "You"-messages can also be investigated without using a classifier. As initially stated, correctly detecting these messages is out of the scope of this thesis. Instead, terms indicative of "I"- and "You"-messages can be counted like "I", "you" but also "myself", "yourself" and others. The relation of word usage in disruptive to constructive posts is a less stark contrast for these terms as can be seen in Table 5.1. However, they are a lot more frequently used: 36.60% of all disruptive and 26.89% of all constructive posts contain at least one of the terms "I" and "you". Additionally considering the terms "me", "my", "your", "myself" and "yourself" increases this to 40.13% and 29.37% respectively. 43.93% of all disruptive and 32.11% of all constructive posts contain at least two of the terms. Therefore, if these terms and especially "I" and "you" were strong indicators for constructiveness or disruptiveness, they would improve the recall noticeably.

In accordance with the assumed effect of "I"- and "You"-messages, "I" appears significantly more often in constructive than in disruptive posts. Likewise, "you" appears more than twice as often in disruptive posts. Surprisingly, the terms "me", "my" and "myself" appear more frequently in disruptive posts although they are rather typical for "I"-messages. A possible explanation could be that AfD discussions benefit from objective posts. These three terms, however, are often used in a subjective context to express the

term	share of words from disruptive posts (‰)	share of words from constructive posts (‰)
fucking	**0.06**	0.00
fuck	**0.06**	0.01
shit	**0.09**	0.01
i	6.40	**10.70**
you	**10.64**	4.52
me	**2.43**	1.20
my	**3.00**	1.68
your	**3.05**	1.25
myself	**0.22**	0.13
yourself	**0.20**	0.10

Table 5.1.: The table shows how commonly a term appears in disruptive or constructive posts. A bold font indicates that the term appears more frequently in that class.

author's personal feelings or thoughts. When they are expressed towards another editor, they can be framed as a personal attack, which would make them disruptive.

The relation of the occurrence frequency of "I" to "you" also seems to be expressive. That is, a constructive post contains on average 2.37 times more "I" than "you". Disruptive ones on the other hand have an average relation of 0.60, i.e. "you" appears nearly twice as often as "I" in a disruptive post.

On average, there is a significantly higher use of the term "I" in constructive posts as questioned in RQ1 and the opposite holds true for RQ2. Thus, the results of this simple data analysis support the research questions when exclusively considering the terms "I" and "you" as indicators for "I"- and "You"-messages. Interestingly, other terms that are also likely to encode "I"- and "You"-messages are not in favour of RQ1. The results of the later classifications will not be able to support RQ1 especially in regard to the use of the term "I". RQ4 raised the question which other terms are characteristic for constructive or disruptive messages. This analysis implies that swear words are a strong indicator for disruptive messages although they are only rarely used.

6. Test Setup

This chapter describes the tests and their analysis. The first section introduces the goals of the later tests. It is thereby explained how our analysis could answer the research questions. Section 6.2 describes how the tests were conducted in general. Furthermore, our own approach is introduced which allows the analysis of a short history of an editor's posts. As the results in Section 7.2 show, the classifiers perform mostly worse with this approach. Finally, Section 6.3 describes our model validation and how the annotated data set was sampled for the tests.

6.1. Goals

The motivation of this thesis was to study the effects of function words as well as "I"- and "You"-messages on the disruptive character of messages in textual discussions. For this, we employed the classifiers presented in Section 5.1 on AfD discussions as our model. In respect to function words, this means comparing the performance of full text classifiers against that of classifiers only considering function words. RQ3 raised the question whether function words suffice for distinguishing disruptive and constructive messages. If the differences between the function words and full text classifiers were negligible, this question would be answered in the affirmative. This would also be the case, if the function words classifiers performed notably worse but still significantly better than a random classifier. Naturally, little difference in the classifiers' performances can only answer the question when the full text classifiers already vastly outperformed a random classifier.

We have not developed elaborate measures to detect "I"- and "You"-messages due to the complexity of this task. Instead, as these concepts are frequently reduced to the usage of "I" and "you", we solely inspect the impact of these terms on the classifiers. If the concept of "I"- and "You"-messages would hold true on AfD discussions, "I" would appear among the most important features indicating that a post did not lead to a block. This complies to RQ2. Likewise, "you" would be among the most important features which indicate a disruptive post, which complies to RQ1. It will be evaluated on the SVM and NB classifier but not on the LM classifier as it returns metrics for n-grams instead of single terms. But more importantly, the LM classifier did not perform significantly better than a random classifier. To determine other terms, which are characteristic for either disruptive or constructive posts and thus to potentially answer RQ4, the heaviest weighted features of the SVM and NB classifier are inspected.

6.2. Test Setup and the Sliding Window Approach

After having chosen a timeframe of 1 day and preprocessing the data, a new balanced annotated data set was sampled from the annotated data set created, containing all 21,213 disruptive posts and equally many, randomly sampled constructive posts. The corpus used for determining the best performing timeframe was not reused as it contained only a subset of elements from both classes. At first, all three classifiers were applied to the full text. In a second run, the classifiers only considered function words. We relied on the list of function words for the English language compiled by Leah Gilner and Franc Morales[1]. It includes pronouns (e.g. "I", "you"), prepositions (e.g. "for", "at"), conjunctions (e.g. "but", "and"), auxiliary verbs (e.g. "can", "will"), determiners (e.g. "her", "his") and quantifiers (e.g. "some", "none").

In addition to those classifications that consider posts individually, tests have been made which consider parts of an editor's post history. Subsequently, the later approach will be called *sliding window* and the original one *independent posts* approach. In this approach, posts by the same author made shortly after each other are merged into a single new post. A merged post is compiled from posts where the oldest and newest ones are not further apart than 1 day. We call this timespan a *window*, which slides over posts and creates the longest merged post using every post as the oldest one once. The time difference of 1 day is identical to our previously determined best performing timeframe because it yielded the best results in the earlier tests using isolated posts. Tests with different window sizes could not be conducted due to the limited scope of this thesis.

The process of merging an editor's posts into a short history of their posts is depicted in Figure 6.1 and can be described as follows: All posts are separated by editor and sorted chronologically, starting with the oldest. The first post of an editor is grouped together with all posts which they authored within 1 day afterwards in a single window. Thus, the time difference between the oldest and newest post in this window is not greater than 1 day. They are then merged keeping their chronological order. Subsequently, the window is shifted to start at the second oldest post and the process is repeated until all posts are processed. Therefore, single posts may appear in multiple merged posts. As a result, the average post becomes longer while the total number of constructive and disruptive posts in test and training data remains the same. In the example given in Figure 6.1, the post with content "`C`" will appear three times in the final data set.

If a post was made within the timeframe of 1 day before the author was blocked, it is assumed to have led to this block. In the example, a block must have happened at or after 2.5 days but before 3 days. Every window containing such a post will also be regarded as having led to a block. Regarding the example, the merged posts with contents "`CD`", "`DE`" and "`E`" will be considered as disruptive. Therefore, the number of posts identified to be disruptive may increase and the number of constructive ones may decrease analogously. The original data set in Figure 6.1 contained two disruptive posts and the one created by our sliding window approach contains three. Likewise, the original annotated data set created from the Wikipedia dumps contained 179 times more

[1] `http://www.sequencepublishing.com/academic.html` – last accessed 10 October 2015, 22:00

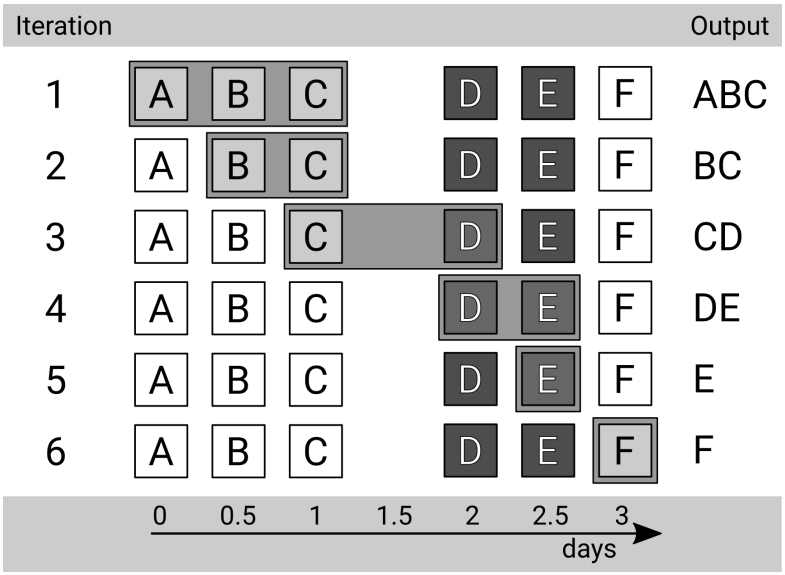

Figure 6.1.: The graphic shows the sliding window algorithm throughout six iterations together with the merged posts that it outputs. Posts are represented by a square with the letter symbolising its content. All posts are assumed to be authored by the same editor. The grey rectangle depicts the current window. A disruptive post is indicated as dark grey square with its content in white font. The window size and timeframe are both set to be 1 day.

constructive than disruptive posts. The data set built with the sliding window approach reduces this, so that there are approximately 162 times more constructive than disruptive posts. Thus, there are still sufficient constructive posts for creating a balanced data set. To have the results somewhat comparable, the same sample size was chosen to be identical to the independent posts approach.

Assuming that blocking has an educational effect, editors might change their behaviour and make only constructive posts afterwards. The algorithm respects this potential change in tone: A window will only slide over a disruptive post when the chronologically next post belongs to the same block. Hence, if a merged post is considered disruptive, at least its chronologically newest post must have been disruptive as well. If the window cannot grow anymore due to a disruptive post being followed by a constructive one or a disruptive post from another block, the algorithm will continue as if this disruptive post was the last post within the 1 day timeframe. Therefore, increasingly shorter windows will be created until the window only spans the last disruptive post. Figure 6.1 illustrates this behaviour in iterations 3 to 5. Afterwards, a new window will be constructed starting at the first post chronologically following this disruptive post. In the example given, this is the post with content "`F`".

Just like the independent posts tests, the sliding window approach was evaluated once using full text classifications and a second time using function word classifications. All tests were evaluated using the metrics presented in Section 5.2.

6.3. Model Validation

All classifiers are learnt and tested using a 10-fold cross-validation. In this validation, the data is being partitioned into ten parts. The classifier is learnt on nine of them and tested on the left out tenth. Thus, training and evaluation data are separate. The process is repeated ten times, so that every partition has been the test corpus once.

We implemented the cross-validation of the LM classifier so that the set of constructive posts and the set of disruptive posts were linearly partitioned. Thereby, we ensured that the data sets which the classifier was learnt and tested on were always balanced. Regarding the SVM and NB classifier, we used stratified sampling for the independent posts approach. This was done to ensure that the training and testing data was balanced in every cross-validation iteration. Different to linear sampling, the data is randomly partitioned while ensuring a balanced class distribution. Therefore, the posts used for training and testing of all three classifiers are identical but the partitioning differs between the LM classifier and the two other classifiers implemented in RapidMiner. RapidMiner's implementation of linear sampling is less suitable for the input data as it would result in a class imbalance for every partition.

However, tests using the sliding window approach were performed using linear sampling for the SVM and NB classifiers. Partitions featuring unbalanced class distributions had to be accepted because the sliding window approach allows posts to appear in multiple merged posts. As stratified sampling creates partitions from randomly selected data, merged posts could appear in the test set that contain posts which were already in the training set. Problems related to the differences in cross-validation are discussed in Section 8.3.

7. Results

In this chapter, the results of our classifiers throughout the various tests are shown. Concrete performance values can be found in Appendix A. For an easier visualisation, the AUC values have been transformed to percentages in all following charts. As decided upon earlier, posts made up to 1 day before a block of their author were considered disruptive for all tests. For less cluttered charts, the names of the metrics have been abbreviated. A plus symbol (+) indicates a performance metric calculated for disruptive and a minus symbol (−) for constructive posts, e.g. positive recall will be labelled "Recall+".

The first two sections present the results of the independent posts and the sliding window approach. After that, the results of applying the independent posts approach to the oldest AfD discussions are shown. This data led to eminently improved results, different to all other tested data sets. For all classifications, the differences in performance of full text and function words classification are illustrated. Moreover, terms which are characteristic for disruptive and constructive posts are discussed.

7.1. Independent Posts Approach

All values used to generate the charts in this section are given in Table A.2. Figure 7.1 shows the performance of the three classifiers considering all words. Overall, the support vector machine performs best, the naïve Bayes classifier ranks second and the language model ranks last. The SVM is outperformed only in negative recall by the NB classifier. However, due to lower negative precision, the NB classifier results in a worse negative F1 score than the SVM. The SVM is significantly better in positive recall and AUC, whereas the NB classifier and the LM classifier repeatedly perform similarly. When looking at the positive and negative F1 scores, it can be seen that all classifiers are better in predicting constructive posts than disruptive ones. It was expected that the SVM would outperform the NB classifier. Yet, we imagined that the language model would perform better. All in all, the three classifiers perform hardly well enough to draw reliable conclusions from the results. F1, AUC and accuracy scores of about 80% and above would be needed.

The full text classifiers perform significantly better than their function words equivalents. Except for negative recall of the SVM and LM classifier as well as the AUC for the NB and LM classifiers, all metrics indicate a worse performance when solely considering function words as visible in Figure 7.2. A decline was expected because the amount of information in a post was massively reduced. Around 59.56% of words in a post have been ignored in the function words classifications. The SVM's performance changes the most whereas the LM classifier's performance remains similar. Nevertheless, the LM classifiers' performances cannot be used to answer RQ3 regarding the importance of

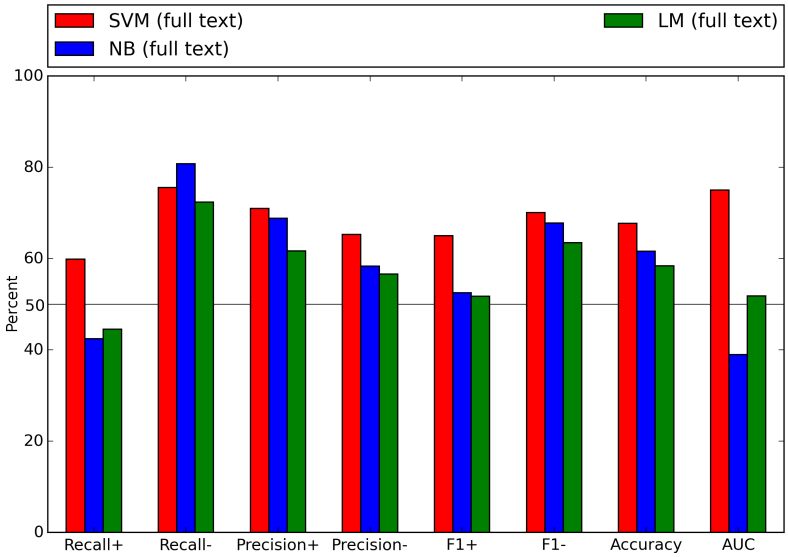

Figure 7.1.: The performance of the full text SVM, NB and LM classifiers using the independent posts approach.

function words. The reason for this is that the full text classification's performance is already too similar to that of a random binary classifier that alternates its predictions with $p = 0.5$ independent of any input. This is visualised in Figure 7.3, which shows the performance differences between the full text and function words LM classifier side by side in relation to such a random classifier. The function words classifier achieved an accuracy of merely 55.73%.

Figure 7.4 contains the five most characteristic terms for constructive and disruptive posts each. The terms were retrieved by extracting the highest and lowest weighted terms according to the full text SVM as it performed the best. Seven out of these ten terms are specific to Wikipedia and AfD discussions. This indicates that the SVM's results are probably not representative for textual discussions in general. "Fancruft" is a Wikipedia term for contents that are only of interest for a small number of avid fans [W27]. "Nom" is short for "nomination" and often used in posts such as "`delete as per nom`" to express agreement towards the deletion nomination. "Wikipedia:Deletion review" is detected as a single term due to our preprocessing step that converts internal Wikipedia links appropriately. Deletion reviews offer Wikipedia users to appeal article deletions if they believe that the deletion was unrightful by Wikipedia standards [W26]. "Small" often appears as an HTML element to markup contents in a smaller font size. At this point, it is unclear why many of these elements have not been removed in the data preprocessing steps. "Redirect" and "keep" are both potential outcomes for the discussed article.

Editors recommending to "keep" an article express that in their opinion the article's quality and relevance are both sufficient and that the article should not be deleted. One explanation for this term being ranked the highest could be that it may be used in

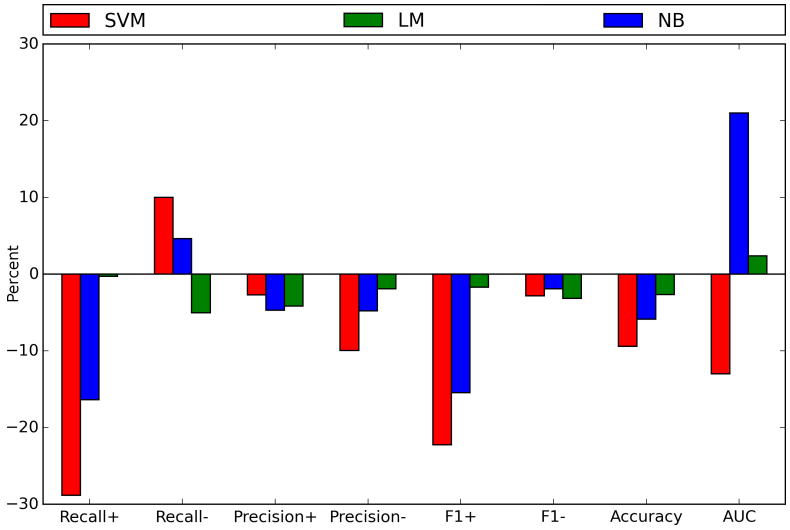

Figure 7.2.: This figure shows the difference in performance between the function words classifiers and their full text equivalents. Positive percentages show a performance improvement over the full text classifier and negative ones a decline. For example, if a classifier considering all words achieved 50% AUC and that considering function words 70%, the function words classifier performed 20% better overall.

contexts where an editor was involved in creating the article and will subjectively defend it. Some editors even create new accounts to act as different editors who argue to keep the article as well. Such accounts are called *sock puppets* on Wikipedia and are highly likely to be blocked once detected [W32]. Besides, manual inspection indicated that disruptive posts containing "keep" often also contained personal attacks. The SVM seems to have optimised for a frequently appearing term, with it being contained by 22.75% of all disruptive and 19.90% of all constructive posts in this data set. On the other hand, the highest weighted swear word "fuck" appears in 0.77% to 0.05% of the posts respectively. In relation to each other, "fuck" appears much likelier in disruptive than in constructive posts than "keep" does. Yet, it was only ranked 17th place.

The results support RQ1 in that "You"-messages seem to be commonly used in disruptive posts. "You" and "your" both appear in the top five of most characteristic terms for disruptive posts. "I" ranks 62nd. For comparison, "I" is the 36,201st heaviest weighted term identifying constructive posts. Regarding these results, "I" is thus not typical for constructive posts which was speculated by RQ2.

As can be seen in Figure 7.5, the SVM yielded the best results when considering only function words. Consequently, we inspected the ten heaviest weighted stemmed terms. They are shown in Figure 7.6. It is difficult to speculate why the function words have been weighted as they are. However, "you" and "your" both being part of the top five terms most characteristic for disruptive posts in this data set indicates support for RQ1. The two terms are common parts of "You"-messages. Moreover, "you" is by far the most

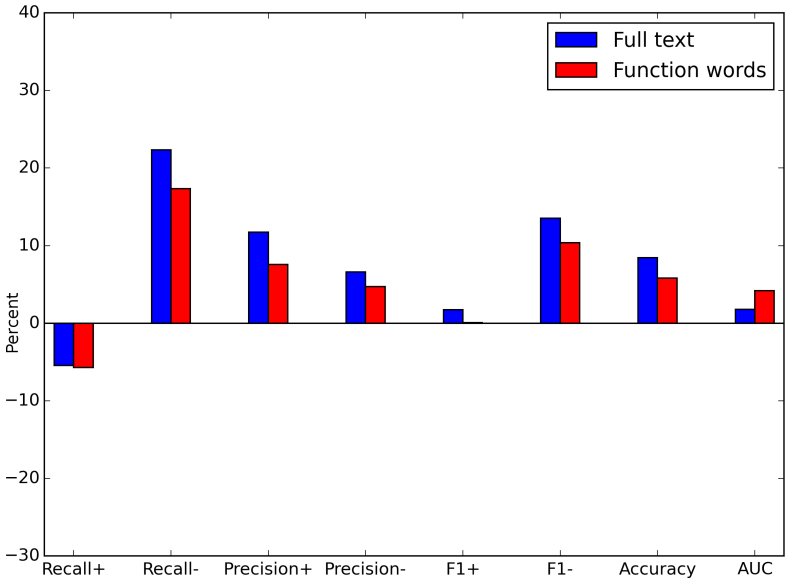

Figure 7.3.: This bar chart shows performance metrics in relation to the performance of a random classifier with $p = 0.5$ for both classes. Depicted is the full text LM classifier in contrast to the function words LM classifier. Values below 0% indicate a performance worse than a random classifier, whereas values above 0% mark a better performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result.

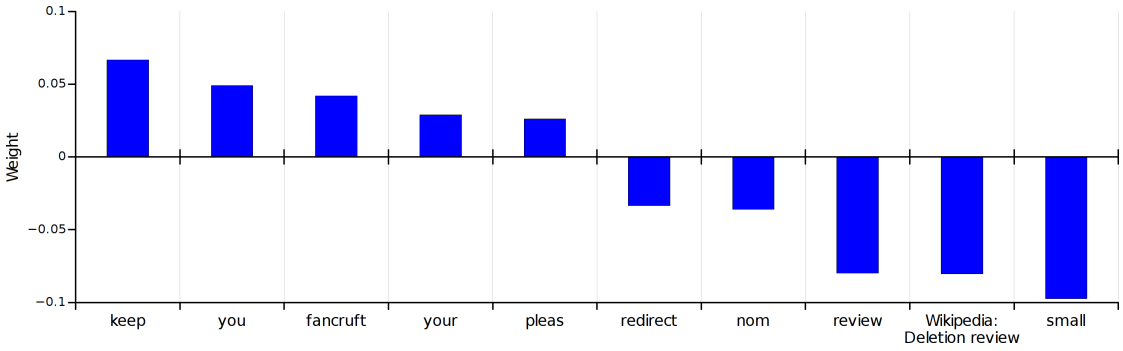

Figure 7.4.: The top five lowest and highest weighted stemmed terms of the full text SVM. Positive weighted terms are characteristic for disruptive and negative weighted for constructive posts. The penultimate word was formatted to contain a colon and a space for better readability.

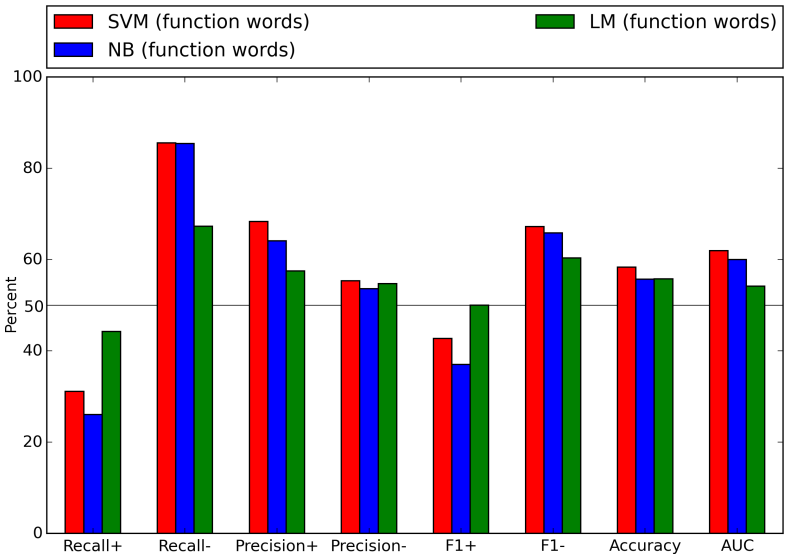

Figure 7.5.: The performance of the SVM, NB and LM classifiers using the independent posts approach while only respecting function words.

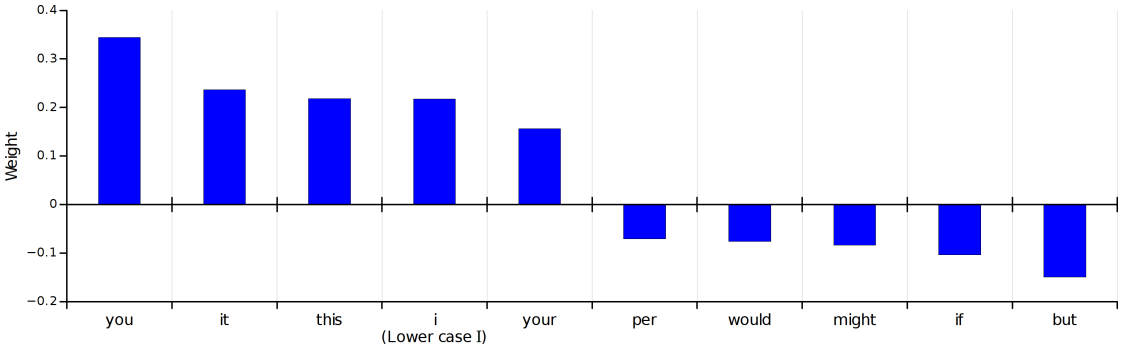

Figure 7.6.: The top five lowest and highest weighted stemmed terms of the function words SVM. Positive weighted terms are characteristic for disruptive and negative weighted for constructive posts.

characteristic term for disruptive posts among all considered 162 function words. In contrast, RQ2 questioned whether "I"-messages would typically be used in constructive messages but "I" was the fourth highest positively weighted term. Hence, RQ2 seems to be negated.

Nonetheless, it is important to take account of the performances when drawing conclusions from the observed results. The function words SVM achieved merely 58.34% accuracy. Thus, RQ1 being supported and RQ2 being negated should be interpreted as a tendency. Also, the nature of AfD discussions might generally favour objective posts and thus bias how "I"- and "You"-messages are perceived. With an accuracy of only 67.73%, the SVM performs significantly better than a random classifier. Yet, its results are not good enough for answering the research questions with certainty. Thus, although "you" and "your" are in the top five of terms characteristic for disruptive posts, RQ1 cannot be safely answered. The same holds true for "I" being identified as typical term contained in disruptive posts, which would negate RQ2. However, both amplify the trends seen in the function words classification. We found no support for function words being sufficient for making good predictions about whether a message is constructive or disruptive as was asked by RQ3. RQ4 addressed whether there are other characteristic terms. No answer can be given that would be applicable to general textual discussions as other terms identified as being characteristic for either of the classes were related to Wikipedia and AfD discussions.

7.2. Sliding Window Approach

Classifying posts independent of each other implies that only the posts within the 1 day timeframe were of disruptive nature. This might not always be the case and the results of that approach were not satisfying either. Thus, the sliding window approach was tested as well. It increased the number of posts in our data set that were assumed to be disruptive.

The number of constructive and disruptive posts was increased compared to the independent posts classifications. This is conditioned by the sliding window algorithm which regards all merged posts containing a blocked post as disruptive. As this approach considers an editor's post history, the data is different and the approach was evaluated on newly sampled data. However, multiple runs of classifications using both the independent posts as well as the sliding window approach showed that the results always remained similar. All values used to generate the charts in this section are given in Table A.3.

Figure 7.7 shows the performance of the classifiers using the sliding window approach. As RapidMiner was unable to calculate the AUC, it had to be left out. When inspecting the accuracy, it becomes clear that the performance of all three classifiers has worsened. The SVM was the formerly best performing classifier but has turned into the worst performing. Again, the NB classifier performs slightly better than the LM classifier. With the independent posts approach, the LM classifier was significantly better in predicting constructive posts as expressed by the F1 scores. This tendency was amplified in the sliding window approach as its positive F1 score is now 22.03% and its negative 66.61%.

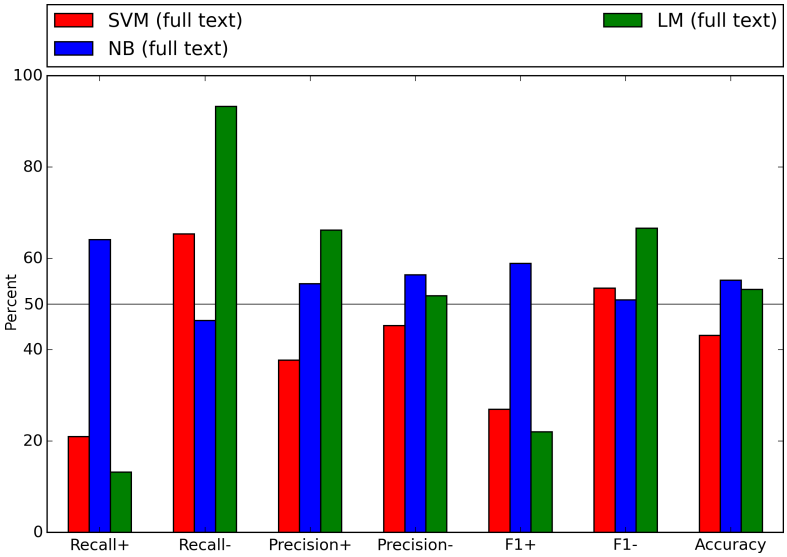

Figure 7.7.: The performance of the full text SVM, NB and LM classifiers using our sliding window approach.

As all classifiers performed similar to or worse than a random classifier, the difference in performance between the full text and function words classification is less distinctive. Surprisingly, the NB classifier forms an exception as all its values improved. Figure 7.8 shows this improvement. With an accuracy of 60.70%, it performed only slightly worse than the full text NB classifier using the independent posts approach which achieved 61.60%. With a positive F1 score of 64.08% and a negative of 56.61%, it predicted disruptive better than the full text NB classifier of the first approach. The latter had positive and negative F1 scores of 52.49% and 67.78% respectively.

Consequently, we extracted the terms most characteristic for constructive and disruptive posts from the function words NB classifier. Figure 7.9 shows the five function words most typical for constructive posts. All these terms have in common that they appear infrequently in the analysed posts. We decided to only depict five terms as the following appear similarly meaningless. On the other hand, the terms indicating disruptive posts are quite expressive. They are given in Figure 7.10. Inspecting the data showed that "anti" was mostly used in contexts of strong disagreement. Moreover, it was used in political and religious topics such as "anti-Serbian", "anti-Muslim" and "anti-Semitic". These topics are known to be prone to controversial discussions on Wikipedia [57], which themselves often lead to blocks. "Against" and "anti" have synonymous meanings and suggest that disagreement and disruptive posts are potentially related. As was asked in RQ1, "You"-messages seem to be an indicator for disruptive posts. "Yourself", "you" and "your" are all likely to be used in "You"-messages and belong to the six function words most characteristic for disruptive posts. Similarly, "myself", "me" and "my" are likely part of "I"-messages. Them being identified to be among the strongest indicators for disruptive posts suggests that RQ2 must be negated. "I" appears as 38th most characteristic term

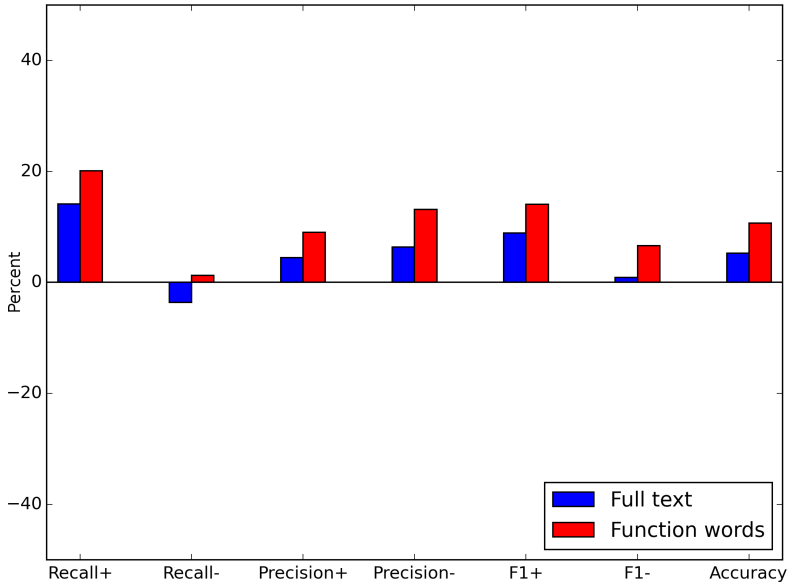

Figure 7.8.: This bar chart shows the performance of the full text NB classifier in contrast to it only factoring in function words when using the sliding window approach. It is set in comparison to the performance of a random classifier with $p = 0.5$. 0% expresses equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result.

for disruptive posts. In total, 162 function words were considered. Regarding the terms from Figure 7.9, it seemed as if their occurrence frequency substantially implied their ranking. This, however, is not true as becomes clear when inspecting the occurrence frequencies of "dare" and "you" for example. "dare", ranking second most characteristic term for disruptive posts, appears merely 93 times in the analysed posts. Conversely, "you" appears 49,997 times and ranks fifth most characteristic term for disruptive posts.

In summary, we saw the sliding window approach perform worse than the independent posts approach. Due to the full text classifiers performing similar to a random classifier, a presentation of characteristic stemmed terms for disruptive and constructive posts was skipped. Despite the low accuracy of 60.70%, the NB classifier yielded better results when considering function words exclusively than when considering all words in a post. We count this as relatively strong support for RQ3 that function words can indeed play an important role in distinguishing constructive and disruptive posts. The observed effects of "I"- and "You"-messages in the independent posts approach were also visible in the function words NB classification. Therefore, our presumptions are reinforced that RQ1 could be affirmed and RQ2 negated. As asked in RQ4, the list of most influential function words for disruptive posts hinted that terms expressing opposition could be an indicator for determining whether a post is disruptive.

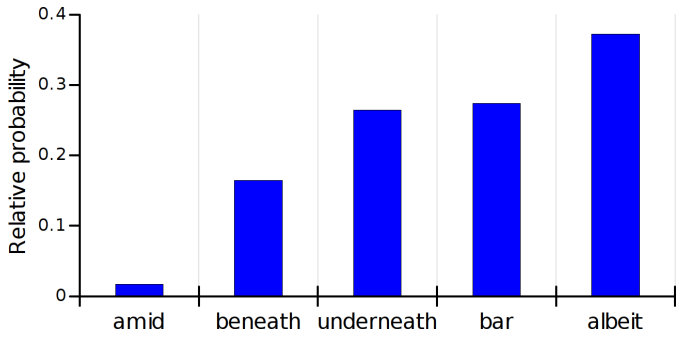

Figure 7.9.: This chart shows a relation of positive probability divided by negative probability for a stemmed term. A low relational probability indicates a term characteristic for constructive posts. Shown are the terms most characteristic for constructive posts according to the results of the function words NB classifier using the sliding window approach.

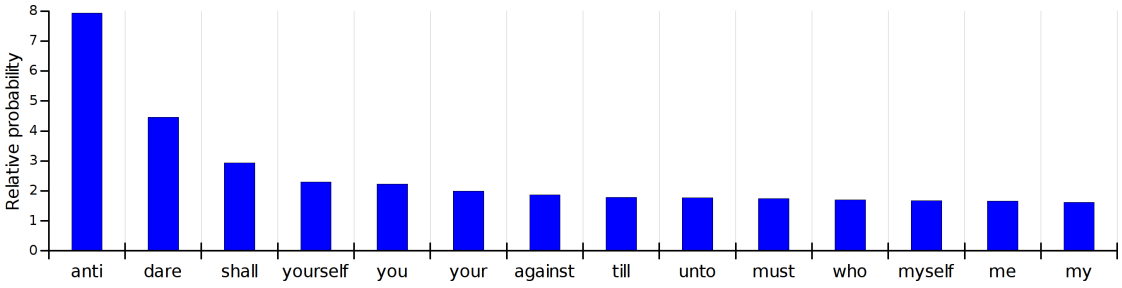

Figure 7.10.: This chart shows a relation of positive probability divided by negative probability for a stemmed term. A high relational probability indicates a term characteristic for disruptive posts. Shown are the terms most characteristic for disruptive posts according to the results of the function words NB classifier using the sliding window approach.

7.3. Analysis of the Oldest Articles for Deletion Discussions

Although we only present the data of a single run each, all tests have been executed multiple times with newly sampled data sets. This confirmed that the observed performances were not the result of a sampling bias. The results always remained comparable but for one exception. When operating on a data set chronologically sampled from the earliest AfD discussions, we found that the classifiers yielded significantly improved performance. Chronologically sampling data from more recent AfD discussions could not reproduce these results. Instead, newer chronologically sampled data was comparable to the performance of our randomly sampled classifications as presented in the previous Sections 7.1 and 7.2. However, due to the amount of existing data, we were not able to test all possible partitions. All values used to generate the charts in this section are given in Table A.4.

As before, all available disruptive posts were considered. Instead of randomly sampling the same number n of constructive posts like in earlier approaches, we used the first n constructive posts in chronological order starting with the oldest. The performance of all classifiers improved using the independent posts approach, cf. Figure 7.11.

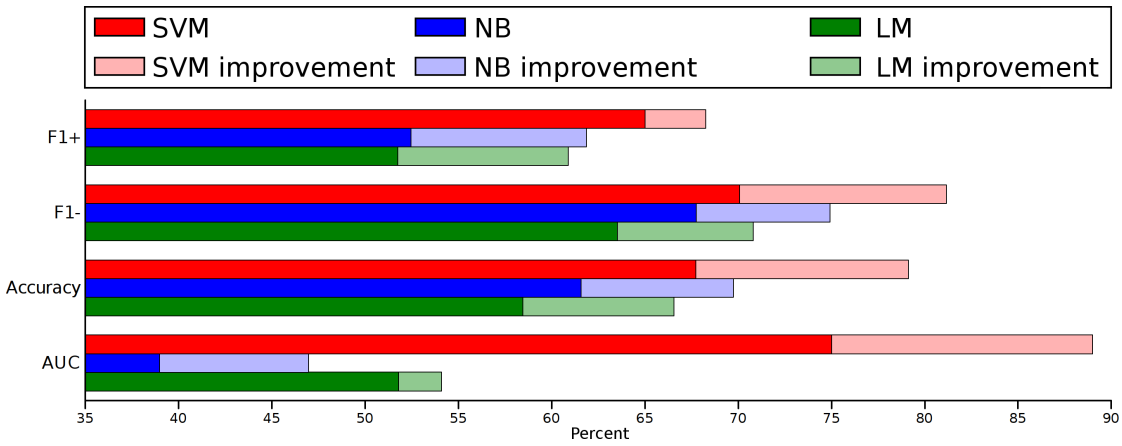

Figure 7.11.: The darker coloured bars show the performance of the independent posts approach using the full text classifiers taken from Figure 7.1. When the analysed data is chronologically sampled from the oldest posts, the performance improves as indicated by the lighter coloured bars. To prevent the chart from becoming unclear, precision and recall have been left out.

Similar to the results observed when using randomly sampled data, the SVM performed best and the language model worst. The SVM using the independent posts approach scored an accuracy of 79.12% and an AUC of 0.890. Moreover, it achieved a positive precision of 87.20% and a negative of 73.93%. Therefore, the highest and lowest weighted

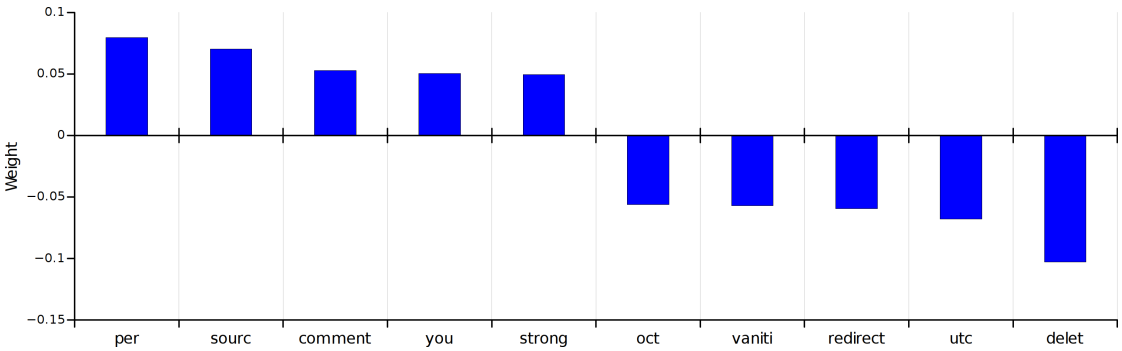

Figure 7.12.: The top five lowest and highest weighted stemmed terms of the full text SVM using the independent posts approach on the chronologically first 19,237 posts. Positive weighted terms are characteristic for disruptive and negative weighted for constructive posts.

terms are likely to be characteristic for constructive and disruptive posts in the context of this special data set. These words are shown in Figure 7.12. Again, most words are specific to AfD discussions. "Per" is another term appearing in the frequently used phrase "`delete per nom`". "Sourc" is a word stem from words such as "source" and is commonly found in posts which highlight that the article in question is lacking trustworthy sources. "Strong" is often used in combination with "keep" or "delete" by editors to express strong belief that the article under discussion should be kept or deleted.

"Oct" is short for October. "Utc" is frequently found in Wikipedia signatures as abbreviation for coordinated universal time [cf. W31]. Although AfD discussions are imaginable in which editors talked about the month October or the coordinated universal time, it is likelier that both words are leftovers from customised signatures. Manual data inspection confirms this assumption. This would then indicate that our attempts to remove user signatures were insufficient for one or multiple users. As a result, the classifier might not have been learnt on the used language itself but has learnt implicit clues that identify one or more disruptive or constructive users. This could partly be a reason why the classifiers yielded notably better results with this data set.

"Vaniti" is a stemmed term derived from the word "vanity". On Wikipedia, vanity was a term to describe contributions that are not of general interest but are made for the purpose of self-promotion [W37]. The term was used 36,830 times until autumn 2004, when it got replaced by the broader behavioural guideline "conflict of interest" [W22]. This guideline covers the potential bias that contributions may contain when their author is in any way related to its content. The term "vanity" never appeared again after 2004.

The heaviest weighted stemmed term indicating constructive posts is "delet". Derived from the verb "delete", it is commonly used by editors to advise for deletion of an article. It could be speculated that words stemmed to "delet" are mostly used in posts that neutrally express to remove an article. Manual inspection supports this presumption as authors of disruptive posts typically advocate to keep an article. However, "per" being the most characteristic stemmed term for disruptive posts contradicts this as it

mostly appears in the phrase "`delete as per nom`". The term "you" is again in the top five of highest weighted words, which indicate disruptive communication. Thus, the associated RQ1 is further supported. The term "I" is the 64th most representative term for disruptive posts. Considering that this data set contained $34,627$ terms, the term is clearly still not typical for constructive posts. As a result thereof, RQ2 is again negated.

When only considering function words, the performance decreases significantly. Figure 7.13 shows the performance of all classifiers. It is slightly better than that of the independent posts approach with randomly sampled data. The overall performance according to accuracy and AUC is again best for the SVM. Figure 7.14 shows the heaviest weighted stemmed terms from the function words SVM. Although the SVM has the lowest positive F1 score, it has the highest positive precision. Therefore, it predicted only few disruptive posts but when it did, it was correct 76.21% of the time. Interestingly, the absolute values of the positive weights is much higher than that of the negative weights. "You", "your" and "I" appear in the five highest weighted terms indicating disruptive posts. As before, this supports "You"-messages being typically used in disruptive posts (RQ1) but negates the use of "I"-messages in constructive posts (RQ2).

In essence, all classifiers perform notably better when analysing the oldest AfD discussions. This is probably due to terms that uniquely appear in this data set like "vanity" as well as not fully removed customised signatures. The results suggest that the signatures belonged to highly active, constructively contributing editors. Nevertheless, "you" has been identified as fourth highest weighted term indicating disruptive posts when considering all words. Together with the term "your", it appeared again among the highest weighted terms characteristic for disruptive posts when only considering function words. RQ1 thus gained further support. The term "I" reappeared among the same terms. Again, RQ2 would have to be negated.

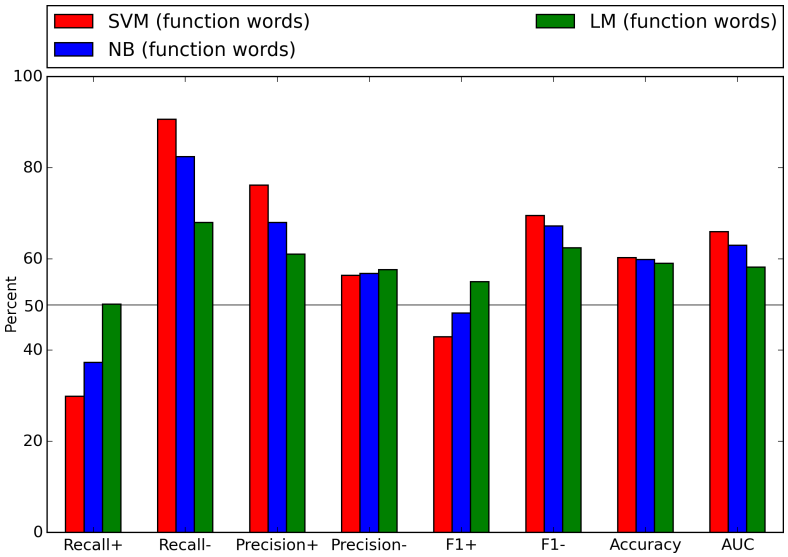

Figure 7.13.: The performance of the function words SVM, NB and LM classifications using the independent posts approach on the chronologically first 19, 237 posts.

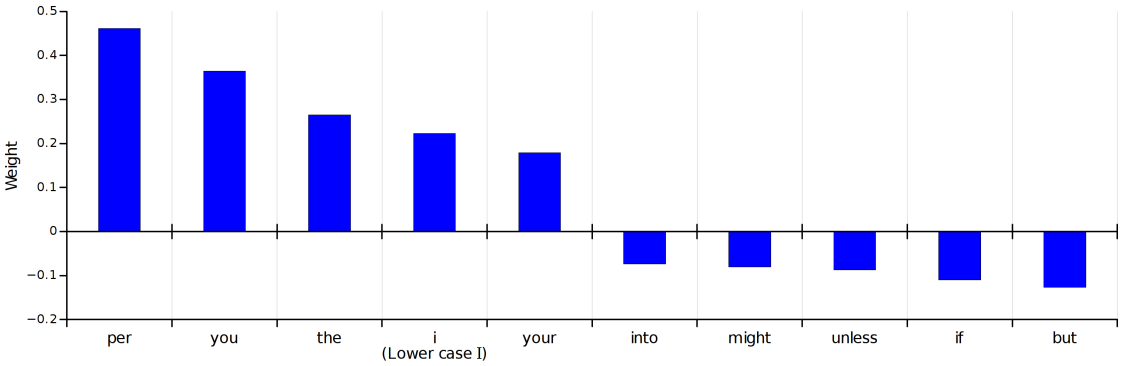

Figure 7.14.: The top five lowest and highest weighted stemmed terms of the function words SVM using the independent posts approach on the chronologically first 19, 237 posts. Positive weighted terms are characteristic for disruptive and negative weighted terms for constructive posts.

8. Design Decisions, Potential Errors and Improvements

Multiple assumptions had to be made throughout this thesis, which could have noticeably influenced the results. This chapter focuses on these systematic errors and design decisions. When possible, it is discussed how they could have influenced the results or how the associated errors could be resolved. Although we put a lot of effort into all steps, the amount and extent of problems potentially impacting the classifications illustrate that the data is not well-suited for the analysis of textual discussions on a word-level.

8.1. Difficulty of Building an Annotated Data Set

A great difficulty was to build an annotated data set by determining which posts may have or have not led to a block. This problem is rooted in the ambiguous information given about issued blocks. Blocks are not associated with any contribution or page. Thus, it is unclear which posts can be labelled disruptive and which ones constructive. The block log even contains cases where the blocking was unrelated to any contribution. For example, an administrator noted in a block log comment that receiving legal threats via email made him issue this very block.

All posts within a fixed timeframe prior to their author being blocked were considered to be disruptive. We determined the best performing timeframe which still contains a great amount of data. However, this cannot ensure that all contributions within the timeframe prior to a block have actually made the administrator act. For example, it is possible that an editor was simultaneously active in multiple AfD discussions before being blocked. In some AfD discussions, they may have been contributing constructively, while they lost their temper in others. Hence, the posts from all their active discussions shortly before their block would wrongfully be labelled disruptive although this is only true for some of them. Our approach cannot detect this behaviour which reduces the correctness of our annotated data set.

Moreover, it is hard to determine relevant blocks based on a free text comment alone while ensuring that the amount of data is not too much reduced. Therefore, the earlier presented blacklisting approach was chosen as a compromise. However, there are reasons for blocks which do not allow to draw conclusions about the blocked editor's post contents. Sock puppetry [W32] was given $109,042$ times as a reason for blocking a registered editor. Thus, 13.62% of blocks on registered editors have been issued due to editors acting as different people, e.g. by creating new accounts or by borrowing the accounts of friends. The contents of posts by sock puppets vary from aggressive comments containing verbal

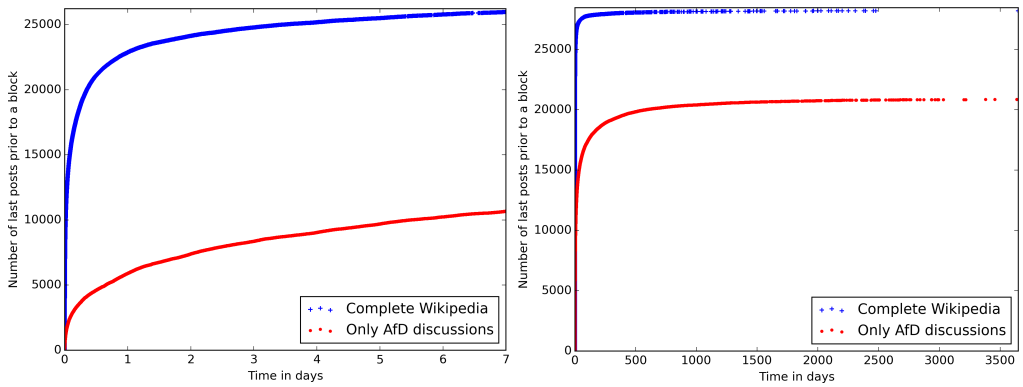

Figure 8.1.: Number of last posts in an AfD discussion and on any Wikipedia page prior to their authors being blocked within a given timeframe in days. Only registered editors who were active in AfD discussions were considered. The left graph shows the first seven days of the right graph in more detail.

attacks to reserved and seemingly objective input. Considering the later, the intention of these posts is to influence others and change their minds while concealing that multiple contributions have been made by the same person disguised as different editors. Their posts should therefore be classified as constructive from a word-level perspective. Yet, when the sock puppets are identified as such, they will be blocked. Regardless of choosing to ignore or consider blocks made due to sock puppets, there will be posts falsely labelled. Future work building a model from similar data should therefore dismiss posts by editors who have at some point been identified as a sock puppet.

It is unclear whether blocks are a good indicator for disruptive communication. They are often only used as ultima ratio in discussions when it is believed that the discussion would benefit from them [W21]. This may frequently not be the case as AfD discussions are mostly only active for seven days before a decision is made [W18]. Additionally, one of Wikipedia's behavioural guideline advises users to always assume contributions to be made in good faith unless there is hard evidence against it [W19]. Figure 8.1 visualises the time difference between the last post of a registered editor and them being blocked when only considering posts made in AfD discussions compared to those made on any Wikipedia page. Only registered editors were considered who created at least one post in an AfD discussion. Blocks were filtered using the blacklisting approach presented in Section 4.4. The total number of posts is higher when considering posts from all Wikipedia pages because $7,363$ registered editors have already been blocked once or multiple times before they ever contributed to an AfD discussion. As can be seen from Figure 8.1, the curve that considered all Wikipedia pages grows much faster. This indicates that many registered editors, who have been active in at least one AfD discussion, have been blocked due to misbehaviour outside of these discussions.

A solution could be to use an annotated data set that was compiled by humans. To our knowledge, there does not exist one for AfD discussions yet. Furthermore, the task is non-trivial as it is located on a post-level without any context. This became especially

clear, when we randomly picked thirty posts and asked four people to classify them. The test cannot be seen as a serious user study but serves as mere impression on the difficulty of the task. This difficulty is expressed in a Fleiss' kappa of 0.18. While only considering two classes (predicting a post to lead to a block or not within 1 day), the low value implies that there was poor agreement among the participants. Moreover, with an average accuracy of 55%, we saw the participants performing worse than the software classifiers in our independent posts approach.

8.2. Difficulty of Preprocessing Posts

This section summarises the difficulties associated with identifying and processing posts. It also shows problems which we had to ignore due to the limited scope of this thesis.

The first discussed difficulty is the identification of posts. Although there are recommendations on how to highlight one's posts within a Wikipedia discussion, these are not always adhered to. Moreover, editors may move or correct text by others or restore older revisions due to vandalism. Both further complicate this task and thus, calculating the differences between one revision and its previous one does not suffice. This problem was solved by using WikiWho, which is the current baseline for this task. Nonetheless, it is still possible that some words do not get attributed to the correct author. We also saw two instances where multiple posts had been identified by WikiWho as one.

It was important to remove non-linguistic and machine generated contents to ensure that only human communication is part of the analysis. With the removal of the wikitext markup, potential stressing of words e.g. via bold font or italics was also removed. We accepted this loss in information as there are no standardised semantics for the use of wikitext markup. Although extending an existing solution for wikitext removal, some fragments still remained. That is for one due to the complexity of wikitext, which also allows the use of HTML and CSS. Likewise, it is related to many posts containing erroneous HTML syntax as well as not containing well-formed HTML. This problem is potentially amplified by errors made by the WikiWho algorithm. For example, let us assume a post contained an italic emphasis "`<i>this</i> is great`". If WikiWho wrongfully associated the first less-than sign to a different post, the wikitext removal would no longer detect an element opening tag. As a result, the word "`i`" would be added twice although it should not be added at all. This is because a closing tag is only removed together with its matching opening tag. Manual inspection showed mostly satisfying results but it has not been systematically evaluated.

To remove content that was automatically generated by algorithms of Wikipedia, signatures and templates had to be detected. Transcluded templates were easily removed but detecting substituted templates is a difficult task. We were able to remove the templates most commonly used in AfD discussions. Templates which were infrequently embedded in AfD discussions are still present in the data which was used for analysis. Due to them being seldom used, their effect should be negligible. If, however, they share many words, their singular effects could add up and influence the results.

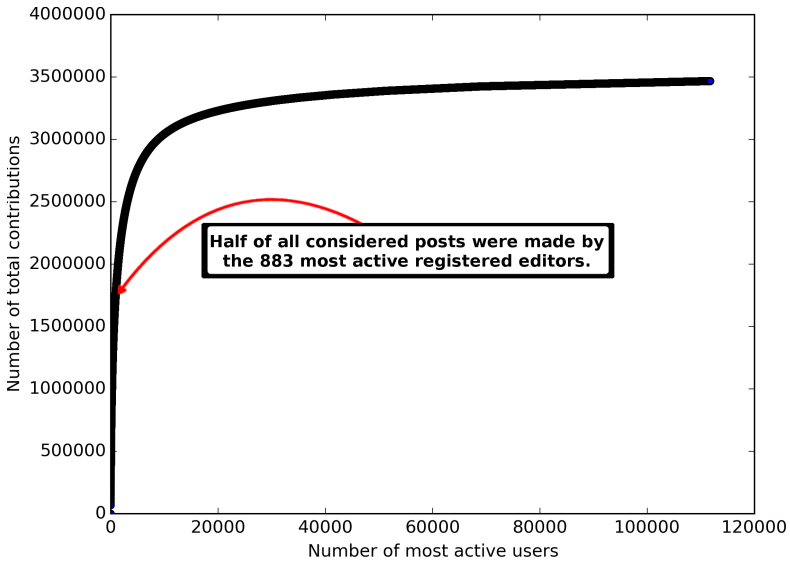

Figure 8.2.: The graph shows the number of total posts made by the number of most active registered editors.

Signatures were removed using simple regular expressions. Yet, they do not match all customised signatures. Ortega et al. [37] have shown that approximately 10% of all Wikipedia editors make 90% of all contributions. Similarly, in AfD discussions around 11.51% of registered editors have made 90% of all posts. Figure 8.2 illustrates the effect which a few registered editors have on the number of posts. A fictitious user called "Steve" might decide to modify his signature to contain "`I am Steve`". If this signature is not fully removed in the preprocessing step and if Steve is a heavily active editor, he can influence the results noticeably. Following this example, if Steve makes many constructive posts, the term "I" might be identified as important feature for constructive posts. Moreover, the classifiers would learn to identify editors instead of language characteristics.

It can be assumed that signatures are mostly customised by very active editors because of multiple reasons. First, not everyone may be aware of the possibility to customise one's signature and more complex customisation requires knowledge in wikitext, HTML and CSS. Very active users are thus likelier to be aware of such extended features of Wikipedia and likewise to be more proficient in Wikipedia's own markup. Second, very active users are probably also more interested in customised signatures for being noted and recognised. Hence, the incomplete signature removal can impact the word distribution. The full text SVM classification using the independent posts approach with the oldest AfD discussions illustrated this. It identified the two terms *oct* and *utc* among the top five most characteristic stemmed terms for disruptive posts. These are assumed to be parts of signatures that have not been fully removed. Manual inspection of posts extracted from more recent AfD discussions also showed that some parts of signatures from highly active registered editors remained.

The further removal of the most common symbols including dashes, periods, commas, and brackets may have impacted our analysis. It was a necessity to exclusively analyse a post's words and to not overcomplicate the task. However, e.g. emotions in text expressed using smilies such as "`:)`" are lost. The use of smilies can be ambiguous and is culturally dependant [38]. Additionally, they can be used in ways where they contradict the actual content of the message e.g. as part of personal attacks like "`you are an idiot :)`". "Netspeak" is also not considered as it is an unstandardised Internet slang. For example, the abbreviated "`b4`" instead of "'before'" will be reduced to "`b`" throughout our preprocessing steps. Smilies and "netspeak" are both more prevalently used with messaging services that have a maximum message length to save characters and to communicate emotions. AfD discussions, however, focus on objective collaboration instead of expressing emotions and feelings. Less than 0.40% of all considered posts, including customised signatures and links, contained "`:)`" or ":-)" before their removal. Therefore, we assume that ignoring smilies has not heavily affected the classification performance.

Due to its complexity, no efforts were made to detect irony or sarcasm. Not only is detecting sarcasm in written text a difficult task for algorithms [cf. e.g. 30, 47] but also for humans [cf. e.g. 1, 17]. Heavy use of sarcasm could have influenced the results as the messages would seem constructive on a word-level but can be disruptive from a semantical perspective.

All words were transformed to lowercase in order to analyse the occurrence of terms while disregarding their capitalisation. Therefore, the first word of a sentence—usually starting with a capital first letter—will not be different from when it had appeared in the middle of the sentence. Words that are written in all capital letters are frequently perceived on the Internet as shouting and its author to being angry. This information is lost as is the differentiation between homographs of different capitalisation. Homographs are words which are spelled identically but have different meanings. For example, "Dick" is a diminutive for the name Richard and will be grouped together with a slang word for the male genitalia which is often used as an insult. Nevertheless, we argue that these cases only form rare exceptions.

To exclusively analyse human communication, posts by users whose name ended in "Bot" were ignored. Posts by bots that do not follow the naming scheme for bots will still be left in the analysed data. However, we estimate these chances low as bots for organisational tasks abide by the naming convention. Additionally, bots built for vandalism purposes are assumed to be unlikely to operate on AfD discussions as the harm done to Wikipedia is less visible there than on regular articles.

8.3. Difficulty of Creating Comparable Results

As mentioned earlier, the classifiers use different implementations of the 10-fold cross-validation. Overall, they consider all given posts for training and testing. However, the partitioning between our own implementation and that of RapidMiner differs. Therefore, the SVM and NB classifier are comparable with each other but not fully with the

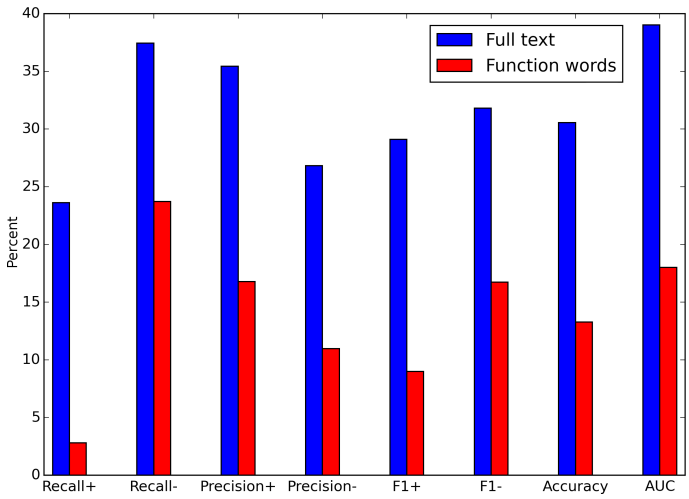

Figure 8.3.: This bar chart shows the performance of the full text SVM in contrast to it only factoring in function words. The classifier used the sliding window approach with stratified sampling. Thus, it erroneously performed well. The values are in relation to a random classifier with 0% expressing equal performance. Table A.5 contains the values which were used to generate this chart as well as the results of the NB classifier. The percentages refer to the overall performance, meaning that 50% equates to a perfect result.

LM classifier. Although very likely, it can hence not be said that a language model classifier does indeed fit the task of distinguishing constructive and disruptive posts in AfD discussions less.

Due to different sampling, the sliding window and independent posts approaches as used by the SVM and NB classifier are not fully comparable either. With the independent posts approach, we use stratified sampling as it is most similar to the sampling used for the LM classifier. However, this was not an option for the sliding window approach or else the classifiers could potentially be tested on data they were partly learnt on. Figure 8.3 visualises the performance of an SVM using the sliding window approach with stratified sampling. It is set in comparison to a random classifier with $p = 0.5$ for each class. Due to this sampling, the SVM performs better than a random classifier in all metrics. As a result thereof, the tests using the sliding window approach were carried out using linear sampling. Hence, the training and testing phases of the sliding window approach used unbalanced data. Five iterations were run with a relation of 4 disruptive to 5 constructive posts. Consequently, the other five were run with a relation of 5 disruptive to 4 constructive posts. We estimate the impact of it on the results to be small but it must be considered when comparing the results.

Although related, the goal of this thesis was not to determine the best performing classifier. Instead, the classifiers were used as a tool for analysing the data and thereby the usage of words in constructive and disruptive posts. Therefore, in spite of the classifiers not being fully comparable, they were still suited for this task.

9. Conclusion

In this thesis, we set out to analyse the language used in textual discussions on a word-level, especially in regard to disruptive messages. We chose to conduct the analysis using Wikipedia AfD discussions because they provide a great amount of textual discussions. Moreover, they feature sufficiently many instances of disruptive messages to perform a large-scale analysis. The effects of "I"- and "You"-messages and function words were of special interest. Using the block log, we built a model that was evaluated using binary classifiers. Thereby, we extracted terms typical for disruptive and constructive posts.

It must be noted that the subsequent answers to the research questions are based on mediocre classification results. Nonetheless, we conducted different tests with different data samples and repeatedly saw similar results. Therefore, we concluded that the questions could be answered. The answers should not be interpreted as definite but rather as a strong tendency.

That is not only due to the classifiers' performances but also because it is unclear how well results from the analysis of AfD discussions can be applied to general textual discussions. Inspecting the heaviest weighted terms mostly showed terms specific to Wikipedia. It is possible that AfD discussions favour objective posts. Thus, the probability for "I"- and "You"-messages being classified as disruptive because of their subjective nature would increase. Furthermore, we identified several difficulties such as the removal of customised signatures and transcluded templates. In sum, they might have impacted the results notably. Finally, a different model might be better at capturing constructive and disruptive textual messages. Our solution for a fully automated creation of an annotated data set by using blocks might be unsuitable for the nature of AfD discussions. All things considered, we can now answer the research questions as follows:

RQ1: Do disruptive messages in textual discussions contain more "You"-messages than constructive ones?

Yes. Throughout all tests, terms that we deemed characteristic for "You"-messages appeared in the top five of terms typical for disruptive posts. Despite the unsatisfying performances of the classifiers, the classifiers determined "you" to be more likely to appear in disruptive posts than over $30,000$ other terms. The results are in accordance with the occurrence frequency of "you" in the complete available data as was shown in Section 5.4.

RQ2: Do constructive messages in textual discussions contain more "I"-messages than disruptive ones?

Probably not. The term "I" was found to be characteristic for disruptive posts in all tests. It was not as highly ranked in the lists of terms typical for disruptive posts as "you" was. Still, "I" was clearly never a typical term for constructive posts throughout the tests. This is in contradiction to the observations made in Section 5.4. In this section, we saw the term "I" appearing a lot more frequently in constructive than in disruptive posts. One possible explanation could be that there might be a few constructive posts which contained the term "I" many times whereas a great number of disruptive posts contained it only once or twice.

RQ3: Is solely considering function words sufficient for determining whether a message is constructive or rather detrimental to a textual discussion?

Maybe. Due to the already mediocre performance of the full text classifiers, the function words classifiers often performed similar to or worse than a random classifier with $p = 0.5$ for both classes. However, the NB classifier using the sliding window approach performed better when only factoring in function words. Consequently, the question can currently neither be affirmed nor negated.

RQ4: Which other words are typical for constructive and which for disruptive messages in textual discussions?

The results make it difficult to answer this question as the most influential terms were mostly specific to AfD discussions or Wikipedia in general. Results from the tests conducted using the oldest AfD discussions suggest that terms expressing opposition and thus disagreement like "anti" and "against" could be characteristic for disruptive posts. Deducing from the results of discussion analysis from related work, terms associated with controversial topics such as religion could indicate disruptive posts as well. Yet, such words were not evident from our results.

Future work should consider building a new model, e.g. by using manually annotated data. It should also take the language used in AfD discussions into account which is different from that used in general textual discussions. Hence, data from more generic discussions on Wikipedia or a completely different data source could be used instead. This could then bring new insights into the terms characteristic for constructive and disruptive messages in textual discussions as well as the correctness of our results.

A. Omitted Data and Graphics from the Test Results

A.1. Effects of Different Timeframes on Classifier Performance

time	recall_+	recall_-	precision_+	precision_-	F1_+	F1_-	accuracy	AUC
13 hours	**49.81**	75.35	66.68	60.26	56.8	66.87	62.58	0.551
1 day	49.29	**76.63**	**67.83**	**60.27**	**56.93**	**67.42**	**62.96**	**0.562**
1.5 days	47.45	76.03	66.46	59.22	55.18	66.52	61.74	0.542
2 days	48.14	75.0	65.85	59.24	55.38	66.11	61.57	0.542
2.5 days	46.57	75.77	65.88	58.72	54.35	66.09	61.17	0.536
3 days	46.49	74.04	64.33	58.12	53.71	65.03	60.26	0.529
4 days	45.35	74.03	63.72	57.59	52.76	64.7	59.69	0.521
5 days	45.13	74.51	64.07	57.64	52.72	64.92	59.82	0.522
6 days	42.95	74.6	63.02	56.7	50.85	64.35	58.77	0.512

Table A.1.: This table shows the average performance of the classifiers with the various timeframes. The values are the arithmetic mean of the classifiers' results. These are an SVM, a NB classifier and an LM classifier. A bold font highlights the best result considering this metric. The plus and minus symbols indicate whether the performance metric was calculated for disruptive (+) or constructive contributions (−).

A.2. Independent Posts Approach

Classifier	Recall_+	Recall_-	Precision_+	Precision_-	F1_+	F1_-	Accuracy	AUC
SVM	59.91	75.56	71.03	65.33	65.00	70.07	67.73	0.750
SVM (FW)	31.11	85.58	68.33	55.40	42.75	67.26	58.34	0.620
NB	42.42	80.78	68.82	58.38	52.49	67.78	61.60	0.390
NB (FW)	26.04	85.43	64.12	53.60	37.04	65.87	55.73	0.600
LM	44.56	72.35	61.71	56.61	51.75	63.52	58.45	0.518
LM (FW)	44.28	67.34	57.55	54.72	50.05	60.38	55.81	0.542

Table A.2.: This table shows the performance of the SVM, NB and LM classifiers using the independent posts approach. Function words classifications are marked with "FW". All others were full text classifications. The plus and minus symbols indicate whether the performance metric was calculated for disruptive (+) or constructive contributions (−).

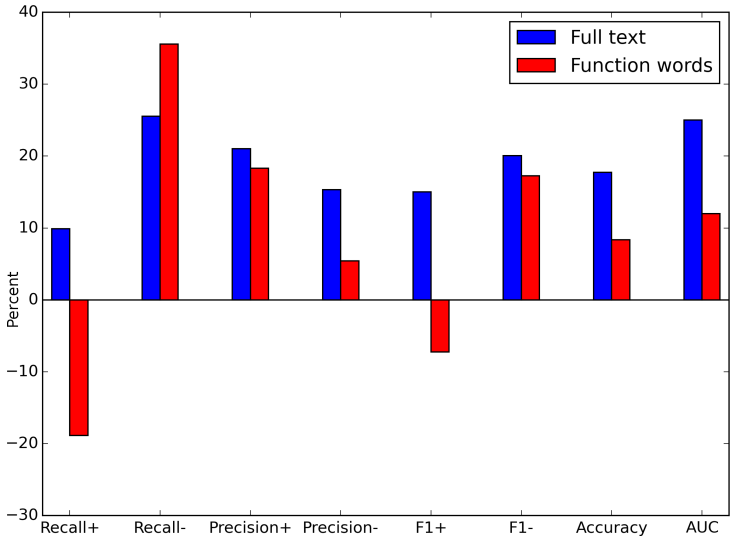

Figure A.1.: This bar chart shows the performance of the full text SVM in contrast to it only factoring in function words when using the independent posts approach. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

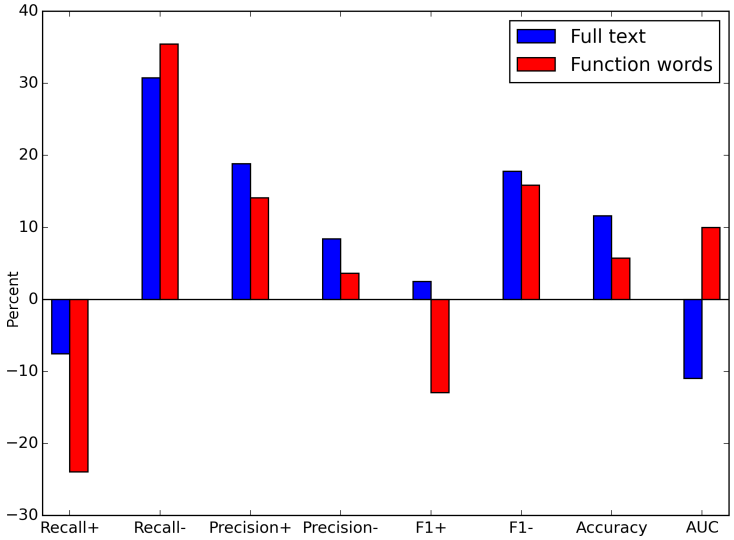

Figure A.2.: This bar chart shows the performance of the full text NB classifier in contrast to it only factoring in function words when using the independent posts approach. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

A.3. Sliding Window Approach Using Linear Sampling

Classifier	Recall_+	Recall_-	Precision_+	Precision_-	F1_+	F1_-	Accuracy	AUC
SVM	20.97	65.39	37.73	45.28	26.96	53.51	43.18	—
SVM (FW)	29.63	37.92	32.31	35.01	30.91	36.41	33.77	—
NB	64.14	46.39	54.47	56.40	58.91	50.91	55.26	—
NB (FW)	70.11	51.28	59.00	63.18	64.08	56.61	60.70	—
LM	13.21	93.26	66.23	51.80	22.03	66.61	53.24	0.497
LM (FW)	4.69	96.88	60.05	50.41	8.70	66.31	50.78	0.559

Table A.3.: This table shows the performance of the SVM, NB and LM classifiers using the sliding window approach with linear sampling. RapidMiner did not return AUC values for the SVM and NB classifier, so they had to be left out. Function words classifications are marked with "FW". All others were full text classifications. The plus and minus symbols indicate whether the performance metric was calculated for disruptive (+) or constructive contributions (−).

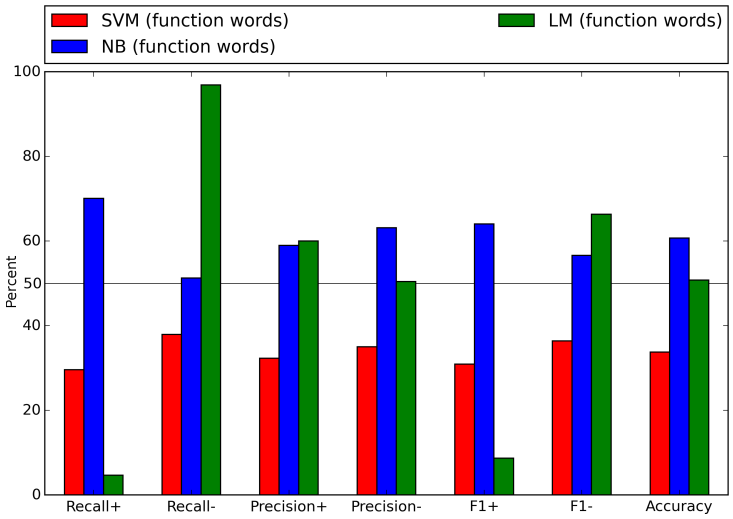

Figure A.3.: The performance of the full text SVM, NB and LM classifiers using the sliding window approach and linear sampling. The visualised values can be found in Table A.3.

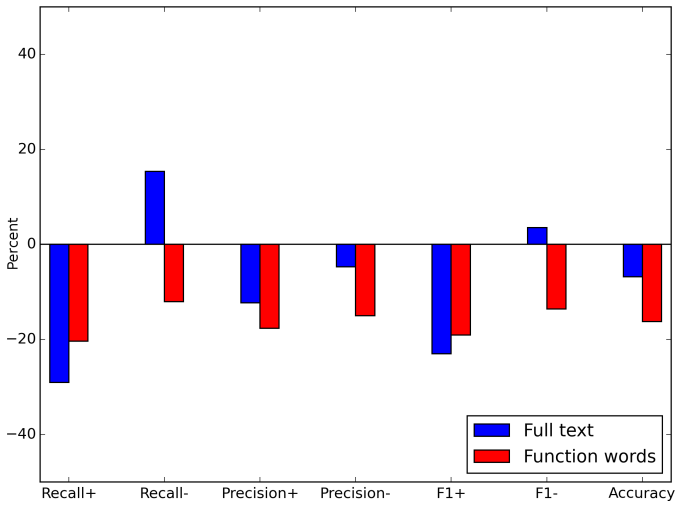

Figure A.4.: This bar chart shows the performance of the full text SVM in contrast to it only factoring in function words when using the sliding window approach and linear sampling. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

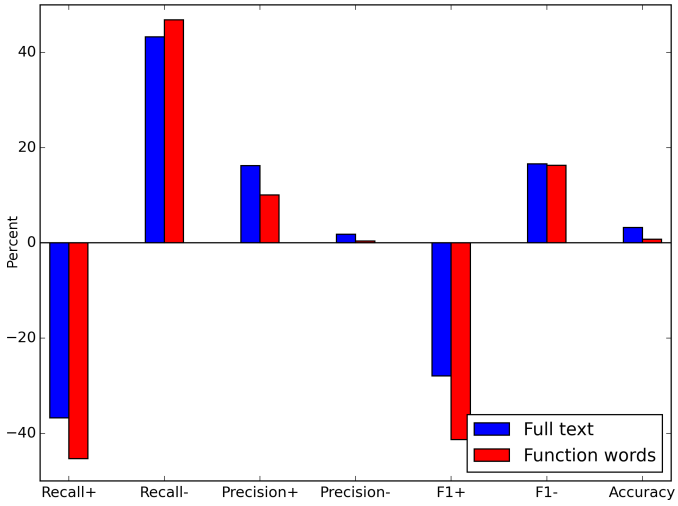

Figure A.5.: This bar chart shows the performance of the full text LM classifier in contrast to it only factoring in function words when using the sliding window approach and linear sampling. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

A.4. Oldest Articles for Deletion Discussions

Classifier	Recall_+	Recall_-	Precision_+	Precision_-	F1_+	F1_-	Accuracy	AUC
SVM	68.26	89.98	87.20	73.93	76.58	81.17	79.12	0.890
SVM (FW)	29.86	90.68	76.21	56.39	42.91	69.54	60.27	0.660
NB	49.10	90.44	83.70	63.99	61.89	74.95	69.77	0.470
NB (FW)	37.30	82.43	67.98	56.80	48.17	67.26	59.86	0.630
LM	52.04	81.08	73.33	62.83	60.88	70.80	66.56	0.541
LM (FW)	50.08	68.04	61.04	57.68	55.02	62.43	59.06	0.582

Table A.4.: This table shows the performance of the SVM, NB and LM classifiers using the independent posts approach on the oldest AfD discussions. Function words classifications are marked with "FW". All others were full text classifications. The plus and minus symbols indicate whether the performance metric was calculated for disruptive (+) or constructive contributions (−).

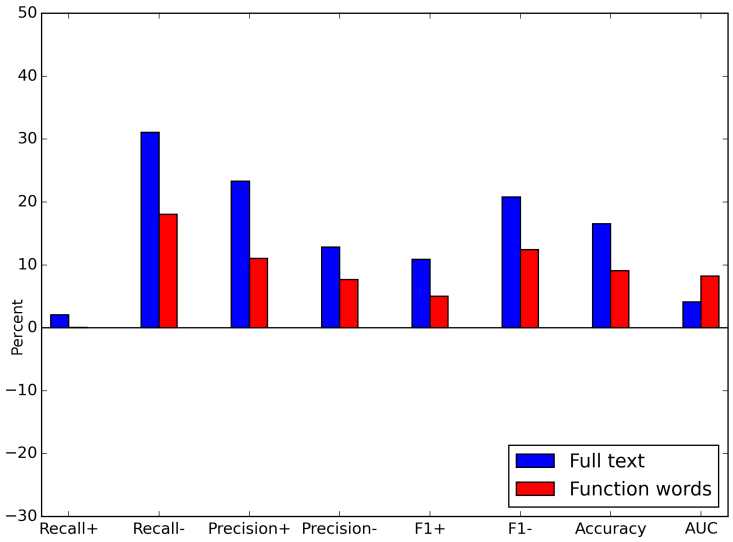

Figure A.6.: This bar chart shows the performance of the full text LM classifier in contrast to it only factoring in function words when using the independent posts approach on the oldest AfD discussions. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

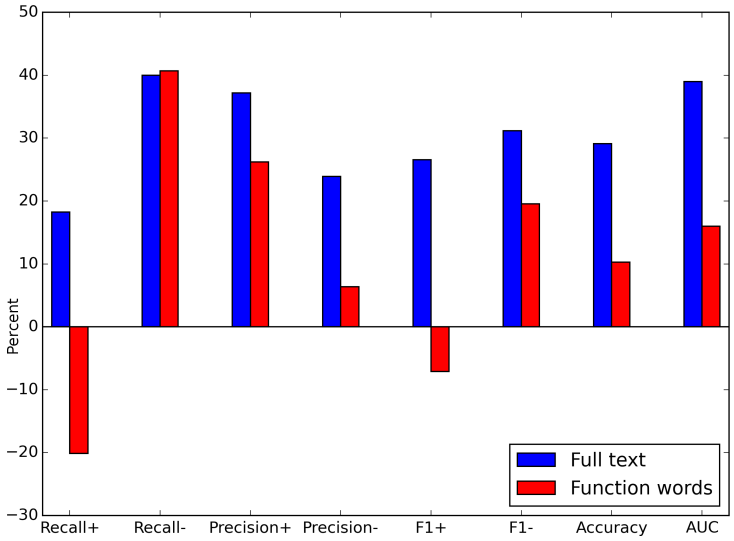

Figure A.7.: This bar chart shows the performance of the full text SVM in contrast to it only factoring in function words when using the independent posts approach on the oldest AfD discussions. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

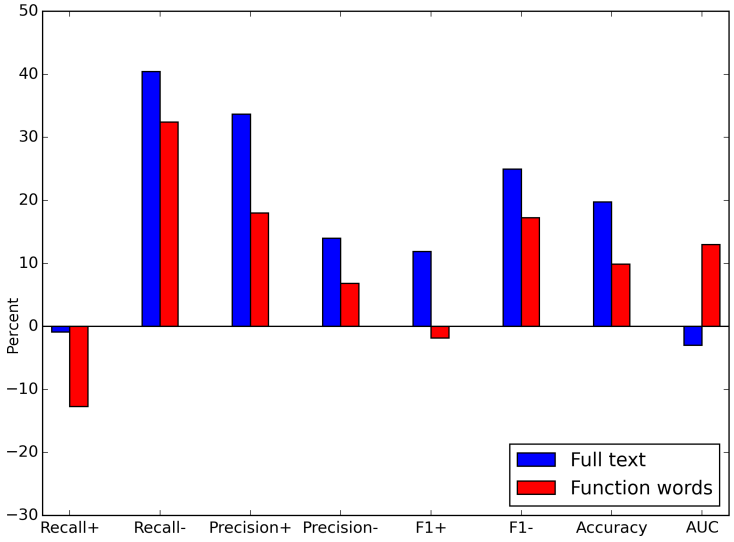

Figure A.8.: This bar chart shows the performance of the full text NB classifier in contrast to it only factoring in function words when using the independent posts approach on the oldest AfD discussions. It is set into relation to the performance of a random classifier with 0% expressing equal performance. The percentages refer to the overall performance, meaning that 50% equates to a perfect result. Table A.2 contains the visualised values.

A.5. Sliding Window Approach Using Stratified Sampling

Classifier	Recall_+	Recall_-	Precision_+	Precision_-	F1_+	F1_-	Accuracy	AUC
SVM	88.96	84.60	85.25	88.46	87.07	86.49	86.78	0.950
SVM (FW)	78.73	50.75	61.52	70.47	69.07	59.01	64.74	0.710
NB	69.57	93.43	91.37	75.43	78.99	83.47	81.50	0.670
NB (FW)	42.27	79.33	67.16	57.88	51.88	66.93	60.80	0.670

Table A.5.: This table shows the performance of the SVM and NB classifier using the sliding window approach with stratified sampling. Function words classifications are marked with "FW". All others were full text classifications. The plus and minus symbols indicate whether the performance metric was calculated for disruptive (+) or constructive contributions (−).

Bibliography

Literature Sources

[1] Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. "How can you say such things?!?: Recognizing disagreement in informal political argument". In: *Proceedings of the Workshop on Languages in Social Media*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 2–11. ISBN: 978-1-932432-96-1. URL: http://dl.acm.org/citation.cfm?id=2021111 (visited on 01/19/2016).

[2] B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features". In: *Computational linguistics and intelligent text processing*. Springer, 2011, pp. 277–288. DOI: 10.1007/978-3-642-19437-5_23.

[3] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. "Annotating social acts: Authority claims and alignment moves in wikipedia talk pages". In: *Proceedings of the Workshop on Languages in Social Media*. LSM '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 48–57. ISBN: 978-1-932432-96-1. URL: http://dl.acm.org/citation.cfm?id=2021116 (visited on 01/11/2016).

[4] Bojan Cestnik. "Estimating probabilities: a crucial task in machine learning." In: *Proceedings of the ninth European conference on artificial intelligence*. Vol. 90. Stockholm, Sweden: IOS Press, 1990, pp. 147–149. URL: http://www.temida.si/~bojan/_resources/Cestnik_Estimating_probabilities.pdf (visited on 10/30/2015).

[5] Eugene Charniak. "Tree-bank grammars". In: *Proceedings of the National Conference on Artificial Intelligence*. 1996, pp. 1031–1036. URL: http://www.aaai.org/Papers/AAAI/1996/AAAI96-153.pdf (visited on 10/26/2015).

[6] Stanley F. Chen and Joshua Goodman. "An empirical study of smoothing techniques for language modeling". In: *Computer Speech & Language* 13.4 (1999), pp. 359–393. DOI: 10.1006/csla.1999.0128.

[7] Si-Chi Chin, W. Nick Street, Padmini Srinivasan, and David Eichmann. "Detecting Wikipedia vandalism with active learning and statistical language models". In: *Proceedings of the 4th workshop on Information credibility*. ACM, 2010, pp. 3–10. URL: http://dl.acm.org/citation.cfm?id=1772942 (visited on 10/20/2015).

[8] Cindy Chung and James W. Pennebaker. "The psychological functions of function words". In: *Social communication* (2007), pp. 343–359.

[9] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/bf00994018.

[10] Luca De Alfaro and Michael Shavlovsky. "Attributing authorship of revisioned content". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 343–354. URL: http://dl.acm.org/citation.cfm?id=2488419 (visited on 08/03/2015).

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. "Predicting Depression via Social Media". In: *ICWSM*. 2013. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124 (visited on 01/08/2016).

[12] William J. Doherty and Robert G. Ryder. "Parent Effectiveness Training (P.E.T.): Criticisms and Caveats". en. In: *Journal of Marital and Family Therapy* 6.4 (Oct. 1980), pp. 409–419. ISSN: 1752-0606. DOI: 10.1111/j.1752-0606.1980.tb01333.x.

[13] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[14] Fabian Flöck and Andriy Rodchenko. "Whose article is it anyway?–Detecting authorship distribution in wikipedia articles over time with WIKIGINI". In: *Online proceedings of the Wikipedia Academy* (2012). URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.360.6261 (visited on 05/26/2015).

[15] R. Stuart Geiger and David Ribes. "The work of sustaining order in wikipedia: the banning of a vandal". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 117–126. URL: http://dl.acm.org/citation.cfm?id=1718941 (visited on 10/20/2015).

[16] Dafydd Gibbon, Roger Moore, and Richard Winski. *Handbook of standards and resources for spoken language systems*. Walter de Gruyter, 1997. ISBN: 978-3-110153-66-8.

[17] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. "Identifying sarcasm in Twitter: a closer look". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 581–586. ISBN: 978-1-932432-88-6. URL: http://dl.acm.org/citation.cfm?id=2002850 (visited on 01/19/2016).

[18] Joshua T. Goodman. "A bit of progress in language modeling". In: *Computer Speech & Language* 15.4 (2001), pp. 403–434. DOI: 10.1006/csla.2001.0174.

[19] Thomas Gordon. *A Theory of Parent Effectiveness*. 1967. URL: http://eric.ed.gov/?id=ED028815 (visited on 03/30/2015).

[20] Thomas Gordon. *Leader Effectiveness Training (L.E.T.) - Proven skills for leading today's business into tomorrow*. Revised and Updated 25th Anniversary Edition. New York, NY, USA: Penguin Putnam Inc., 2001.

[21] Thomas Gordon. *Parent Effectiveness Training: The Proven Program for Raising Responsible Children*. Harmony Books, 2000.

[22] Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. "Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 83–88. ISBN: 978-1-932432-88-6. URL: http://dl.acm.org/citation.cfm?id=2002755 (visited on 10/26/2015).

[23] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. "What's with the attitude?: identifying sentences with attitude in online discussions". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1245–1255. URL: http://dl.acm.org/citation.cfm?id=1870779 (visited on 01/11/2016).

[24] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998. DOI: 10.1007/BFb0026683.

[25] Gerald C. Kane. "A multimethod study of information quality in wiki collaboration". In: *ACM Transactions on Management Information Systems (TMIS)* 2.1 (2011), 4:1–4:16. DOI: 10.1145/1929916.1929920.

[26] Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. "Multinomial naive bayes for text categorization revisited". In: *AI 2004: Advances in Artificial Intelligence*. Springer, 2005, pp. 488–499. DOI: 10.1007/978-3-540-30549-1_43.

[27] Edward S. Kubany, David C. Richard, Gordon B. Bauer, and Miles Y. Muraoka. "Verbalized anger and accusatory "you" messages as cues for anger and antagonism among adolescents". In: *Adolescence* 27.107 (1992), p. 505. URL: https://www.ncbi.nlm.nih.gov/pubmed/1414562 (visited on 10/20/2015).

[28] J. Richard Landis and Gary G. Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174. URL: http://www.jstor.org/stable/2529310 (visited on 01/21/2016).

[29] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. "When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. 2011. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2764 (visited on 06/30/2015).

[30] C. C. Liebrecht, F. A. Kunneman, and A. P. J. van den Bosch. "The perfect solution for detecting sarcasm in tweets #not". In: *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. New Brunswick, NJ, USA: Association for Computational Linguistics, 2013. URL: http://hdl.handle.net/2066/112949 (visited on 01/19/2016).

[31] Andrew McCallum, Kamal Nigam, and others. "A comparison of event models for naive bayes text classification". In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. Citeseer, 1998, pp. 41–48. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.9324 (visited on 10/30/2015).

[32] Albert Mehrabian. *Silent messages: Implicit communication of emotions and attitudes*. 2nd Edition. Belmont, CA, USA: Wadsworth, Inc., 1981. ISBN: 0-534-00910-7.

[33] Santiago M. Mola-Velasco. "Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals: Lab Report for PAN at CLEF 2010". In: *CoRR* (2012). URL: http://arxiv.org/abs/1210.5560 (visited on 10/24/2015).

[34] Christoph Thomas Müller, Willi Hager, and Elke Heise. "Zur Effektivität des Gordon-Eltern-Trainings (PET) — eine Meta-Evaluation". de. In: *Gruppendynamik und Organisationsberatung* 32.3 (Sept. 2001), pp. 339–364. ISSN: 1618-7849, 1862-2615. DOI: 10.1007/s11612-001-0034-7.

[35] Mark EJ Newman. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary physics* 46.5 (2005), pp. 323–351. DOI: 10.1080/00107510500052444.

[36] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. "Lying words: Predicting deception from linguistic styles". In: *Personality and social psychology bulletin* 29.5 (2003), pp. 665–675. DOI: 10.1177/0146167203029005010.

[37] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. "On the inequality of contributions to Wikipedia". In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. IEEE, 2008, pp. 304–304. DOI: 10.1109/HICSS.2008.333.

[38] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. "Emoticon Style: Interpreting Differences in Emoticons Across Cultures." In: *ICWSM*. 2013. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6132 (visited on 12/03/2015).

[39] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. "Augmenting naive bayes classifiers with statistical language models". In: *Information Retrieval* 7.3-4 (2004), pp. 317–345. DOI: 10.1023/B:INRT.0000011209.19643.e2.

[40] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. *Linguistic Inquiry and Wort Count LIWC2007*. Operator's manual. Austin, TX, USA, 2007. URL: http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual.pdf (visited on 05/11/2011).

[41] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. *The Development and Psychometric Properties of LIWC2015*. Austin, TX, USA, 2015. DOI: 10.15781/T29G6Z.

[42] Rene Pickhardt, Thomas Gottron, Martin Körner, Paul Georg Wagner, Till Speicher, and Steffen Staab. "A Generalized Language Model as the Combination of Skipped n-grams and Modified Kneser-Ney Smoothing". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014. URL: http://arxiv.org/abs/1404.3377 (visited on 10/29/2015).

[43] Martin Potthast, Benno Stein, and Robert Gerling. "Automatic vandalism detection in Wikipedia". In: *Advances in Information Retrieval*. Springer, 2008, pp. 663–668. DOI: 10.1007/978-3-540-78646-7_75.

[44] Trisha Prabhu. *Google Science Fair 2014 – Rethink: An Effective Way to Prevent Cyberbullying*. 2014. URL: https://www.googlesciencefair.com/projects/en/2014/f4b320cc1cedf92035dab51903bdd95a846ae7de6869ac40c909525efe7c79db (visited on 12/05/2015).

[45] Russell F. Proctor II and James R. Wilcox. "An exploratory analysis of responses to owned messages in interpersonal communication". In: *ETC: A Review of General Semantics* 50.2 (July 1993), pp. 201–220. ISSN: 0014-164X. URL: http://www.jstor.org/stable/42577446 (visited on 10/20/2015).

[46] Jason D. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, and others. "Tackling the poor assumptions of naive bayes text classifiers". In: *Proceedings of the Twentieth International Conference on Machine Learning 07/2003*. Vol. 3. Washington, D.C., USA): AAAI Press, 2003, pp. 616–623. URL: http://www.aaai.org/Papers/ICML/2003/ICML03-081.pdf (visited on 10/30/2015).

[47] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. "Sarcasm as Contrast between a Positive Sentiment and Negative Situation." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 2013, pp. 704–714. URL: http://www.anthology.aclweb.org/D/D13/D13-1066.pdf (visited on 01/19/2016).

[48] Rachel A. Simmons, Peter C. Gordon, and Dianne L. Chambless. "Pronouns in Marital Interaction What Do "You" and "I" Say About Marital Health?" In: *Psychological science* 16.12 (2005), pp. 932–936. DOI: 10.1111/j.1467-9280.2005.01639.x.

[49] Richard B. Slatcher, Simine Vazire, and James W. Pennebaker. "Am "I" more important than "we"? Couples' word use in instant messages". In: *Personal Relationships* 15.4 (2008), pp. 407–424. DOI: 10.1111/j.1475-6811.2008.00207.x.

[50] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4 (2009), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

[51] Róbert Sumi, Taha Yasseri, András Rung, András Kornai, and János Kertész. "Characterization and prediction of Wikipedia edit wars". In: *Proceedings of the ACM WebSci'11*. Koblenz, Germany, June 2011, pp. 1–3. DOI: 10.1371/journal.pone.0071226.

[52] Lu Wang and Claire Cardie. "A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 2. Association for Computer Linguistics, 2014, pp. 693–699. URL: http://www.cs.cornell.edu/~luwang/papers/ACL2014a.pdf (visited on 01/11/2016).

[53] Walter Weintraub. *Verbal behavior: Adaptation and psychopathology*. New York, NY, USA: Springer Publishing Company, 1981.

[54] Andrew G. West, Sampath Kannan, and Insup Lee. "Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata?" In: *Proceedings of the Third European Workshop on System Security*. ACM, 2010, pp. 22–28. DOI: 10.1145/1752046.1752050.

[55] Yiming Yang and Xin Liu. "A re-examination of text categorization methods". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 42–49. URL: http://dl.acm.org/citation.cfm?id=312647 (visited on 12/12/2015).

[56] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. "Dynamics of conflicts in Wikipedia". In: *PLoS ONE* 7.6 (2012). DOI: 10.1371/journal.pone.0038869.

[57] Taha Yasseri, Anselm Spoerri, Mark Graham, and Janos Kertesz. *The Most Controversial Topics in Wikipedia: A Multilingual and Geographical Analysis*. SSRN Scholarly Paper ID 2269392. Rochester, NY, USA: Social Science Research Network, May 2013. DOI: 10.2139/ssrn.2269392.

Wikipedia Sources

[W1] Wikipedia. *Block log*. Last accessed 16 December 2015, 21:00. 2015. URL: https://en.wikipedia.org/wiki/Special:Log/block.

[W2] Wikipedia. *Category:Articles for deletion templates*. Last accessed 09 August 2015, 14:30. 2013. URL: https://en.wikipedia.org/w/index.php?title=Category:Articles_for_deletion_templates&oldid=547048916.

[W3] Wikipedia. *Category:Wikipedia templates*. Last accessed 09 August 2015, 14:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Category:Wikipedia_templates&oldid=664336614.

[W4] Wikipedia. *Deletionism and inclusionism in Wikipedia*. Last accessed 06 January, 15:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Deletionism_and_inclusionism_in_Wikipedia&oldid=695878369.

[W5] Wikipedia. *Help:Cascading Style Sheets*. Last accessed 14 January, 23:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Cascading_Style_Sheets&oldid=675129002.

[W6] Wikipedia. *Help:HTML in Wikitext*. Last accessed 14 January, 23:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:HTML_in_wikitext&oldid=696338086.

[W7] Wikipedia. *Help:Log*. Last accessed 19 November 2015, 11:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Log&oldid=665919361.

[W8] Wikipedia. *Help:Substitution*. Last accessed 15 January, 2:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Substitution&oldid=697256288.

[W9] Wikipedia. *Help:Template*. Last accessed 05 January, 23:45. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Template&oldid=697252556.

[W10] Wikipedia. *Help:Transclusion*. Last accessed 15 January, 2:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Transclusion&oldid=693549756.

[W11] Wikipedia. *Help:Using talk pages*. Last accessed 18 October 2015, 18:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Using_talk_pages&oldid=684033526\#Indentation.

[W12] Wikipedia. *Help:Wiki markup*. Last accessed 18 October 2015, 14:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Wiki_markup&oldid=685878440.

[W13] Wikipedia. *Liancourt Rocks*. Last accessed 20 October 2015, 22:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Liancourt_Rocks&oldid=686395569.

[W14] Wikipedia. *User talk:SineBot*. Last accessed 18 October, 23:45. 2015. URL: https://en.wikipedia.org/w/index.php?title=User_talk:SineBot&oldid=684188193.

[W15] Wikipedia. *User:Dcoetzee/Willy on Wheels:A Case Study*. Last accessed 15 January, 16:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=User:Dcoetzee/Willy_on_Wheels:A_Case_Study&oldid=676082726.

[W16] Wikipedia. *Wiki markup*. Last accessed 16 December 2015, 20:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wiki_markup&oldid=694764453.

[W17] Wikipedia. *Wikipedia:Articles for deletion is not a war zone*. Last accessed 06 January, 13:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Articles_for_deletion_is_not_a_war_zone&oldid=682504188.

[W18] Wikipedia. *Wikipedia:Articles for deletion*. Last accessed 26 October 2015, 22:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Articles_for_deletion&oldid=686421060.

[W19] Wikipedia. *Wikipedia:Assume good faith*. Last accessed 09 January, 00:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Assume_good_faith&oldid=695906558.

[W20] Wikipedia. *Wikipedia:Block on demand*. Last accessed 13 July 2015, 23:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Block_on_demand&oldid=654197040.

[W21] Wikipedia. *Wikipedia:Blocking policy*. Last accessed 18 October 2015, 09:15. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Blocking_policy&oldid=684928844.

[W22] Wikipedia. *Wikipedia:Conflict of interest*. Last accessed 17 December, 14:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Conflict_of_interest&oldid=694580160.

[W23] Wikipedia. *Wikipedia:Criteria for speedy deletion*. Last accessed 08 January, 22:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Criteria_for_speedy_deletion&oldid=696555278.

[W24] Wikipedia. *Wikipedia:Deletion policy*. Last accessed 26 October, 21:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Deletion_policy&oldid=686722333.

[W25] Wikipedia. *Wikipedia:Deletion process*. Last accessed 23 October 2015 16:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Deletion_process&oldid=683862238.

[W26] Wikipedia. *Wikipedia:Deletion review*. Last accessed 23 January, 20:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Deletion_review&oldid=688090253.

[W27] Wikipedia. *Wikipedia:Fancruft*. Last accessed 13. December 2015, 02:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Fancruft&oldid=694448311.

[W28] Wikipedia. *Wikipedia:No personal attacks*. Last accessed 16 December, 17:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:No_personal_attacks&oldid=692387719.

[W29] Wikipedia. *Wikipedia:Notability*. Last accessed 13 January, 15:00. 2016. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Notability&oldid=699257564.

[W30] Wikipedia. *Wikipedia:Shortcut*. Last accessed 18 November, 21:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Shortcut&oldid=690440021.

[W31] Wikipedia. *Wikipedia:Signatures*. Last accessed 18 October 2015, 18:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Signatures&oldid=684326102.

[W32] Wikipedia. *Wikipedia:Sock puppetry*. Last accessed 21 January, 16:00. 2016. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Sock_puppetry&oldid=699414647.

[W33] Wikipedia. *Wikipedia:Substitution*. Last accessed 15 January, 2:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Substitution&oldid=690854151.

[W34] Wikipedia. *Wikipedia:Username policy*. Last accessed 19 October, 00:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Username_policy&oldid=686393464.

[W35] Wikipedia. *Wikipedia:Using talk pages*. Last accessed 09 January, 00:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Help:Using_talk_pages&oldid=698901153.

[W36] Wikipedia. *Wikipedia:Vandalism*. Last accessed 20 October 2015, 21:30. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Vandalism&oldid=685909877.

[W37] Wikipedia. *Wikipedia:Wikipedia is not a vanity press*. Last accessed 17 December, 14:00. 2004. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_is_not_a_vanity_press&oldid=5617164.

[W38] Wikipedia. *Wikipedia:Wikipedians*. Last accessed 19 October 2015, 17:00. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=685522409.