Computational Linguistics in the Netherlands Journal 6 (2016) 3-20 Submitted 05/2016; Published 12/2016

The Automated Detection of Racist Discourse in Dutch Social Media

Stéphan Tulkens STEPHAN.TULKENS@UANTWERPEN.BE
Lisa Hilte LISA.HILTE@UANTWERPEN.BE
Elise Lodewyckx ELISE.LODEWYCKX@STUDENT.UANTWERPEN.BE
Ben Verhoeven BEN.VERHOEVEN@UANTWERPEN.BE
Walter Daelemans WALTER.DAELEMANS@UANTWERPEN.BE

CLiPS, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium

Abstract

We present two experiments on the automated detection of racist discourse in Dutch social media. In both experiments, multiple classifiers are trained on the same training set. This training set consists of Dutch posts retrieved from two public Belgian social media pages which are likely to attract racist reactions. The posts were labeled as racist or non-racist by multiple annotators, who reached an acceptable agreement score. The different classification models all use the Support Vector Machine algorithm, but use different (sets of) linguistic features, which can be lexical, stylistic or dictionary-based. In the first experiment, the models are evaluated on a test set containing unseen comments retrieved from the same pages as the training set (and thus also skewed towards racism). In the second experiment, the same models from Experiment 1 are tested on an alternative test set, containing more neutral comments, retrieved from the social media page of a Belgian newspaper. In both experiments, the best performing model relies on a dictionary containing different word categories specifically related to racist discourse. It reaches an F-score of 0.47 (exp. 1) and 0.40 (exp. 2) for the racist class and ROC Area Under Curve scores of 0.64 (exp. 1) and 0.73 (exp. 2). The dictionaries, code, and the procedure for requesting the corpus are available at: `https://github.com/clips/hades`.

1. Introduction

The problem of racism in social media has received a lot of attention recently,[1] and the companies behind several popular social networks, such as Twitter and Facebook, have been called upon to improve the detection of, and stiffen the penalties regarding, the racist remarks or the harassment of people for cultural or ethnic background.[2] However, due to the difficulty of discouraging users from engaging in this kind of behavior, most of it continues unabated.

Of course, there are both national and international efforts to ban this kind of behavior from social networks and public life in general. In Belgium, *Unia*, the *Interfederal Centre for Equal Opportunities*, tries to bring attention to racism and discrimination and to provide a way to easily file complaints,[3] which can then be used as evidence in judicial trials. However, the center estimates that racism is often not reported and thus remains invisible (*Discriminatie en Diversiteit: tijd voor een interfederaal actieplan tegen racisme* 2013), as victims of racism often feel that reporting offenses such as slurs and remarks will not solve anything. Additionally, racism on social media can cross international borders, thereby making reporting racism to a national authority rather moot. In this regard, a computational tool that automatically detects racist discourse could be very useful

[1] The Mirror, 18 September 2015: `http://mirror.co.uk/news/technology-science/technology/twitters-racism-epidemic-67-million-6461583` (retrieved on 11/05/2016)

[2] Huffington Post, 3 March 2016: `http://huffingtonpost.com/sydney-latimer/online-activists-question_b_9372588.html` (retrieved on 11/05/2016)

[3] Unia homepage: `http://unia.be/en` (retrieved on 14/09/2016)

as it could provide websites with a way to automatically detect racist comments without people having to report them, thus removing this barrier of entry. Hence, such a tool also allows for a more fine-grained analysis of racist discourse in social media, providing moderators with an overview of the span of such racist posts, regardless of whether they were reported or not. As such, we envision such a tool not as a standalone application which removes comments without any supervision, but as an early warning system, which can serve to alleviate the burden of moderators, or which can help entities such as Unia diagnose racist posts in a timely manner.

In this paper, we present an initial attempt to automatically detect racist language in Dutch social media comments. Our model uses a Support Vector Machine (SVM), which is a supervised machine learning algorithm, to automatically classify posts as either racist or non-racist in a binary classification task. To obtain a gold standard for classification, we manually annotated social media comments which were retrieved from two public social media pages likely to attract racist reactions.

Because this data set is inherently skewed towards racism, we also perform a second experiment using a more neutral test set of comments, retrieved from posts on a newspaper's social media page, which were again manually annotated by the same annotators. In the current research, we examine whether it is possible to accurately predict if a post contains racist language, and, furthermore, whether adding stylistic features to standard lexical features will increase the generality of our model.

Regarding the structure of the paper, we first discuss previous research on racist discourse and its linguistic features, as well as on possible approaches to detecting it. Additionally, we discuss multiple definitions and conceptualizations of racism, and describe how our own definition, which is used in the annotation guidelines and the classification experiments, follows from these conceptualizations (Section 2). Next, we describe the setup, i.e. the data collection, annotation and methodology, of the first experiment, as well as the results (Section 3). Afterwards, we discuss the setup and results of the second experiment (Section 4). Finally, we provide an extensive comparison and evaluation of both experiments (Section 5).

2. Related Research

2.1 Theoretical Definitions of Racism

In order to detect racist discourse, we first need to find an adequate definition of racism. As we want our definition to be implementable, descriptive definitions of racism, of which many have been proposed (see Reisigl and Wodak (2005), Chapter 1, for a thorough overview) do not suffice. Rather, we need a definition which allows us to decide whether a given utterance belongs to the class of racist utterances or not. This is problematic, especially given that racism is bound to the individual; what is racist to some, will probably not be racist to all.

A first option is to turn towards legislation and use the judicial definition of racism in an automated classification procedure. This, however, is not sufficient for our purposes: the Belgian anti-racism law, for example, forbids discrimination, violence and crime based on physical qualities (like skin color), nationality or ethnicity, but does not mention textual insults based on these qualities.[4] Hence, this definition is inadequate, since it does not include racist utterances one would generally find on social media; very few utterances that people might perceive as racist are actually punishable by law, as only utterances which explicitly encourage violence and hate are illegal. While there may be countries in which the anti-discrimination laws are well-suited to be implemented as a classifier, we took the Belgian law as an example because this is the country in which our system would be used.

Because of the difficulty of accurately defining racism, we choose to adopt a simple definition which includes all negative utterances, negative generalizations and insults concerning ethnicity, nationality, religion and culture. In this, we follow Bonilla-Silva, who contends that racism is not

[4] http://unia.be/en/law-recommendations/legislation/discrimination-lexicon (retrieved on 14/09/2016)

limited to physical or ethnic qualities, but can also include social and cultural aspects, as racism as an ideology has moved towards so-called *color-blindness*, which implies the use of "cultural rather than biological explanation of minorities' inferior standing and performance in labor and educational markets"(2002, p.42). It is, in other words, no longer "appropriate" to claim that some people are inferior due to their skin color or genetic heritage. This is adequately summarized by Reisigl and Wodak in the following quote: "it is an undeniable fact for geneticists and biologists that the concept of 'race', with reference to human beings, has nothing to do with biological reality. From a social functional point of view, 'race' is a social construction"(2005, p.2). Because it is no longer possible to give an explanation of the qualities of others on the basis of genetics and biology, racists now look for alternative explanations, such as culture or religion.

Quasthoff proposes another conceptualization of racism, which involves a typology of utterances, all related to the idea of stereotyping, which is an element of common knowledge that is directed towards a certain social group (1989). Examples of (positive) stereotypes are that "all Germans work hard" or "Italians eat pizza". According to Quasthoff, these stereotypes are used to solidify relations within the group to which the speaker belongs (the in-group), and simultaneously to emphasize the otherness of the people that are targeted by the stereotype (the out-group). Hence, all racist utterances can be seen as a form of stereotyping. An interesting note is that even positive generalizations such as the aforementioned remark about Germans can be seen as a form of stereotyping. While we do not deny that there are stereotypes or even racist attitudes at work when making positive prejudiced statements, our research focuses on finding *negative* racist utterances specifically.

Partially based on the work of Quasthoff, Van Dijk describes the importance of creating a divide between an *Us and Them* in racist discourse (2002). Racist utterances often involve a "semantic move with a positive part about Us and a negative part about Them" (Van Dijk 2002, p.150). Using such constructions, one emphasizes - either deliberately or subconsciously - a divide between groups of people. More so than Quasthoff, Van Dijk emphasizes the role of social constructions in the creation of racist ideas (1993). A similar point is made by Coupland, who mentions the concepts of the 'Other' and of 'Othering', which is the "process of representing an individual or a social group *to render them* distant, alien or deviant."(2010, p.244) (emphasis in original). In this sense, linguistic representations, e.g. names or stereotypical descriptions, are often used to *homogenize* groups of people. Overt racist language can be seen as the most extreme form of this procedure, as racist slurs "have their pragmatic effect by forcing an addressed individual into a social group designation, then pejorating that group, by adding explicitly negative attributes or by invoking generally tabooed group labels"(Coupland 2010, p.251). Note that Coupland also includes other characteristics, such as gender (sexism) and age (ageism) as targets for 'Othering', which fall outside the scope of the current project.

2.2 Features of Racist Discourse

Concerning the linguistic content of racist utterances, several authors report markers of racist discourse: Van Dijk (2002) reports that the number of available topics is greatly restricted when talking about foreigners. This is confirmed by Orrù (2015), who, in his qualitative study of Italian Facebook posts, showed that the chosen topics are typically related to migration, crime and economy, even though the pages from which these posts were mined do not have these topics as their focus. In these posts, metaphors such as *waves* or *waterfalls* are often used when talking about immigrants, to denote their supposed number and destructive power. Additionally, immigrants are often referred to in economical terms, as burdens or costs. Reisigl and Wodak (2005) give a similar list, basing themselves on Böke (1997), which includes *thermostatics*, e.g. (economical) pressure, and *animals*, e.g. parasites or rats. Note how all these features fall into the broad category of stereotyping and Othering, as discussed above. By framing, for example, immigrants as lice in our collective hair, we simultaneously tighten the bonds of our in-group, and distance ourselves from the out-group.

Given the extensive theoretical debates on racism and racist discourse, it is surprising that there have been very little efforts towards the automated classification of racist content. Although there has been work on, for example, offensive language on social networks (Djuric et al. 2015, Chen et al. 2012, Nobata et al. 2016), to our knowledge there has been no work that attempts to present a quantification which specifically targets the problem of racism in social media.

Although not working on social media specifically, Greevy and Smeaton also attempt to classify racist utterances using an SVM and several classes of linguistic features (2004b, 2004a). Concerning lexical features, the authors note that racist utterances contain specific words and phrases, like "our own kind" and "white civilization", significantly more often than neutral texts. Second, racist discourse is characterized by a higher rate of certain word classes, like imperatives and adjectives and a higher noun-adjective ratio. In addition, a more frequent use of modals and adverbs is reported, which is linked to the higher frequency of truth claims in racist utterances. Racist utterances can be partially justified by the speaker by casting them as scientific truths instead of opinions (Greevy and Smeaton 2004a, Greevy and Smeaton 2004b).

Additionally, pronoun use is reported as an important feature in the detection of racist language, something which is also noted by Orrù (2015) and Van Dijk (2002). It is important to note that the work of Greevy and Smeaton was performed on a corpus of 1500 *web pages*, 500 of which were racist, 500 of which were anti-racist, and 500 of which were neutral. As such, they do not suffer from the imbalance of racist versus normal comments that is inherent to social media, and have larger amounts of data per instance. The anti-racist corpus balances the racist data by using the constructions of racist discourse in a non-racist way, e.g. by quoting racial slurs to combat or condemn racism itself.

Most similar to our current study is the work of Nobata et al. (2016), who use a feature set similar to ours, including content-, dictionary-, style-based and distributional features. In contrast to our work, they operate on a very large data set of approximately 2 million messages from Yahoo forums, which are not necessarily limited to racist language (but do include racist messages). A combination of all features worked best, achieving an F-score of 0.78, but character n-grams by themselves reached an F-score of 0.77, showing that the additional features only improved the scoring marginally. The style-based features performed worst, followed by the dictionaries, which provides an interesting comparison to our work.

A different view on the detection of more latent forms of racism, conceptualized as ideological discourse,[5] is found in Pollak et al. (2011). In this article, two corpora of newspaper articles about the 2007 Kenyan elections, one consisting of articles from local newspapers and one consisting of articles from Western newspapers, were compared using both quantitative and qualitative approaches. The idea is that differences in word use between these corpora reveal something about the assumptions and ideologies that govern the actions of the journalists reporting the election. An interesting finding was that Western journalists tend to use words which refer to ethnic groups, e.g. *kikuyu* and *tribe*, more often than the local journalists, showing that these categories are most likely a result of an ideology, or implicit ideas regarding the political structure of Kenya. In terms of features, word unigrams, bigrams and trigrams were all tested separately using several decision tree algorithms. Similarly to this work, we also attempt to target more implicit forms of racism, which might only become clear after contrastive analysis. In contrast to Pollak et al., however, we hope to achieve some of these goals solely through automated analysis.

3. Experiment 1

As a first experiment, we attempt the detection of racist discourse in social media comments retrieved from pages which were known for attracting racist remarks. Despite the perceived prevalence of racist comments on social media pages, their actual frequency in normal social media interactions is quite

[5] Note that this use of the word 'ideology' is different from the way it is used by Van Dijk, who sees ideology as a shared collective belief, and is based on work by Verschueren (1999).

low, which is a problem for frequency-based classification approaches such as the one we implement. Therefore, we mined specific social media pages, which are more skewed towards racist comments. In order to find these pages, we contacted Unia, who supplied us with a list of problematic social media pages. From this list, we selected two pages; the first one served as a community hub of a prominent anti-Islamic organization, while the second one was used to post articles by a well-known right-wing organization.[6]

3.1 Data Collection

To collect data, we used `Pattern` (De Smedt and Daelemans 2012) to scrape the 100 most recent posts from both social media pages, and then extracted all reactions to these posts. To highlight why we extract comments and not the posts themselves, we will sketch the modus operandi of these pages. Instead of posting racist content themselves, which is in violation of most social media networks' terms of service, the pages post articles from news websites which concern, for example, immigration. These news stories then attract comments by the people who subscribe to the page, and may contain racist remarks. The posts of the page itself therefore do not contain racist remarks, but instead serve as a catalyst for the posting of derogatory comments.

Extracting the first 100 posts resulted in 5759 extracted comments: 4880 from the first page and 879 from the second one. The second page attracted a lot less comments per post, likely because it posted more frequently. In addition to this, the organization behind the first page had been figuring prominently in the news at the time of extraction, which could explain the divide in frequency of comments. We note that the second page has more subscribers than the first one, which can therefore not be an explanation for the smaller amount of comments on the second page. The corpus was annotated by two annotators, A and B, who were both students of comparable age and background. When A and B did not agree, a third annotator, C, functioned as a tiebreaker in order to obtain gold-standard labels.

To obtain an independent test corpus we followed the same procedure, albeit at a different point in time. We mined the first 500 and first 116 comments from the first and second page, respectively, which makes the proportion of comments that were retrieved from the pages more or less identical to the proportions in the training corpus. Furthermore, this makes the size of our test set about 10% of the training set, which is a standard proportion in classification tasks.

3.2 Annotations

The comments were annotated with three different labels: 'racist', 'non-racist' and 'invalid'.

The 'racist' label describes comments that contain negative utterances or insults about someone's ethnicity, nationality, religion or culture. This label also includes utterances which equate, for example, an ethnic group to an extremist group, as well as extreme generalizations. The following examples were classified as racist:

1. Het zijn precies de vreemden die de haat of het racisme opwekken bij de autochtonen.
   It is the foreigners that elicit hate and racism from natives.

2. Kan je niets aan doen dat je behoort tot het ras dat nog minder verstand en gevoelens heeft in uw hersenen dan het stinkend gat van een VARKEN ! :-p
   You cannot help the fact that you belong to the race that has less intellect and sense in their brains than the smelly behind of a PIG ! :-p

3. Wil weer eens lukken dat wij met het vuilste krapuul zitten, ik verschiet er zelfs niet van!
   Once again we have to put up with the filthiest scum, it doesn't even surprise me anymore!

[6] Due to the sensitive nature of our data, we have chosen not to publicly disclose the source pages. Detailed information on the data set can, however, be requested by contacting the authors.

The label 'invalid' was used for comments that were written in languages other than Dutch, or that did not contain any textual information, i.e. comments that solely consisted of pictures or links. Before automatic classification, we excluded these from both our training and test set.

The final label, 'non-racist', was the default label. If a comment was valid, but could not be considered racist according to our definition, this was the label we used.

We concede that this annotation procedure is loose and open to interpretation, but, as noted above, there is no clear way to create a hard decision boundary to classify racist texts that still does justice to the phenomenon. Second, we consider this an interesting part of the experiment; if the annotators are able to obtain relatively high agreement despite this definition, it means that there is some implicit understanding regarding the nature of racist discourse.

Note that this annotation scheme favors short posts, as the annotators were instructed to classify a post as racist if any part of it was racist. Consequently, posts containing multiple utterances might be annotated as racist even if only a part of the comment was actually racist. In practice, the comments we gathered from the pages were relatively short, with an average post length of 7.9 words, and a median post length of 6 words, which indicates a skewed distribution. Furthermore, the average number of sentences in a post was 2.19 and the median number of sentences was 1, also indicating a skewed distribution.

We calculated inter-annotator agreement using the Kappa score (κ) (Cohen 1968) and simple pairwise agreement. Note that these scores for annotators A and B were calculated before annotator C intervened as a tiebreaker. On the training corpus, the agreement score was $\kappa = 0.60$. Annotator A used the racist tag much less often than annotator B. 79% of the comments that A annotated as racist were also annotated as racist by B, which shows that, even though B was much more inclined to call utterances racist, A and B still shared a common ground regarding their definition of racism. Examining the comments in detail, we found that the difference can largely be explained by sensitivity to insults and generalizations.

The annotation scheme used for the test set was identical to the one for the training set. A difference in the annotation process was that C, who previously performed the tiebreak, also annotated a part of the posts. This was done to assess the degree to which the tiebreaker agreed with the annotators, something which was not possible in the setup used for the training data.[7]

To compute inter-annotator agreement, the first 25% of comments on each page, i.e. 125 comments for the first page and 30 comments for the second one, were annotated by all three annotators. The remaining comments were equally divided among the annotators. The agreement was $\kappa = 0.54$ (pairwise average), which is lower than the score on the training data. The reason for the lower agreement was that annotator C often did not agree with A and B. Because the pattern of mismatches between the annotators is quite regular, we will now discuss some of the annotations in detail:

4. we kunnen niet iedereen hier binnen laten want dat betekend [*sic*] het einde van de europese beschaving
   We cannot let everyone in because that will mean the end of European civilization

5. Eigen volk gaat voor, want die vuile manieren van de EU moeten wij vanaf. Geen EU en geen VN. Waardeloos en tegen onze mensen. (eigen volk.)
   Put our own people first, because we need to get rid of the foul manners of the EU. No EU nor UN. Useless and against our people. (own folk.)

6. Burgemeester Termont is voor de zwartzakken die kiezen voor hem
   Mayor Termont supports the black sacks, as they vote for him

Annotator C used the 'racist' tag more often, which is probably due to the fact that he consistently annotated overt ideological statements related to immigration as 'racist', while the other

[7] The slight difference in annotator setup between the training and test data is not ideal. Unfortunately, in this stage of the research, we can no longer re-annotate our data.

	# Training Comments	# Test Comments
Non-racist	4500 (83%)	443 (73%)
Racist	924 (17%)	164 (27%)
Total	5424	607

Table 1: Gold standard corpus sizes Experiment 1.

annotators did not. The three examples mentioned above are utterances that C classified as 'racist', but A and B classified as 'not racist'. Other than a merely personal difference in sensitivity to insults and generalizations, the cause of these consistent differences in annotations might also be cultural, as C is from the southern part of the Netherlands, whereas A and B are native to the northern part of Belgium. Some terms are simply misannotated by C because they are Flemish vernacular expressions. For example, *zwartzak* [lit. black sack], from sentence 6, superficially looks like a derogatory term for a person of color, but does not usually carry this meaning, as it is a slang word for someone who collaborated with the German occupying forces in the Second World War. While this could still be classified as being racist, the point is that C only registered this as a slang word based on skin color, and not a cultural or political term. Finally, it is improbable that the cause of these mismatches is annotator training, as A and B did not discuss their annotations during the task. In addition to this, C functioned as a tiebreaker in the first data set, and thus arguably had as much experience with the nature of the training material as the other annotators.

The gold standard of the training and test corpus can be found in Table 1.

3.3 Method

3.3.1 Preprocessing

In terms of preprocessing, the text was tokenized and Part of Speech (POS) tagged using the Dutch tokenizer and tagger from `Pattern` (De Smedt and Daelemans 2012), which resulted in lists of tokens which are appropriate for lexical processing. For character-level features, i.e. character n-grams, we used the untokenized text. Following this, we scaled all features to values between 0 and 1 using `scikit-learn` (Pedregosa et al. 2011).

3.3.2 Featurization

Following the broad array of features mentioned and used in both experimental and more theoretical work (see Section 2), we experiment with a large variety of both lexical and stylistic features. By combining these, we hope to capture more implicit forms of racism, e.g. the use of the pronouns 'we' and 'they', as well as explicit racism, e.g. the use of derogatory terms. In the following section, we have grouped our features into three categories: lexical, stylistic and dictionary-based features. Because of the difficulty of categorizing character n-grams, which can capture both lexical and stylistic features, we include these in a separate category. Because the size of the feature sets varied considerably, we summarize the number of features for each category in Table 2.

Lexical features Concerning lexical features, we include lowercased word n-grams, which were shown to be good features for detecting racist utterances in previous research (Greevy and Smeaton 2004b). Through cross-validation, we determined that word trigrams obtained the highest performance.

Stylistic features In terms of stylistic features, we implement average word length, average sentence length (both in words and in characters), and vocabulary richness, i.e. the number of word

Feature class	Number of features
All Features	136,525
Sets	
Only Lexical	124,765
Only Style-based	167
Character trigrams	11,509
Dictionaries	
Racism-related	16
LIWC	68
Combined	84
Baseline	
Word Unigram	14,443

Table 2: The number of features for each feature class, on the training data

types used.[8] Other stylistic features taken into account are POS tag frequencies and punctuation mark frequencies.

Spelling errors and typographic errors could be suitable in this classification task: according to Bonilla-Silva (2002), racist discourse is more linguistically incoherent, and might therefore contain more spelling mistakes or idiosyncratic language use. We did not implement an explicit measure of spelling errors, instead using character n-grams to capture these phenomena. We did not explicitly remove non-alphabetical characters for the purposes of character-based modeling. While removing these characters might be effective to combat masking, i.e. the practice of deliberately inserting punctuation between characters to avoid word-based detection, we did not encounter any attempts at masking in our current dataset.

Finally, we look for differences in the use of three features that are specific to chat language, which is quite similar to the language used in the social media comments. These features include the use of emoticons, flooding, which is the subsequent and deliberate repetition of the same character, and a frequency-based measure of capitalization. In terms of flooding, we include the absolute, relative, and average counts of both character and punctuation flooding.

Note that we do not assume that stylistic features by themselves will be appropriate for the detection of racism. It is ultimately the *meaning* of a sentence, as implied by the words in that sentence, which carries the racist message. Nevertheless, stylistic properties such as exclamation marks, indicating outrage, or emoticon use, indicating sarcasm, might be good indicators for racism.

Dictionaries To capture lexical items which could be indicative of racism, we also use a dictionary-based approach, using the best-performing dictionary from Tulkens et al. (2016b). As shown in the previous work, a dictionary-based approach is more suitable than the analysis of individual word frequencies, as many racist terms in our corpus are neologisms or hapaxes. In effect, a dictionary groups words that are similar to one another, alleviating some of the problems regarding the variety in terminology encountered in the corpus.

The dictionary construction procedure is explained in detail in Tulkens et al. (2016b). As such, we will only give a summary. First, the annotators manually created the dictionary by retrieving possibly racist and more neutral terms from the training data during annotation. This was done before the test set was collected. Because the dictionary terms were extracted from the training data, there is a risk of overfitting, as the racist terms that are used in the test set might not coincide with the terms from the dictionary. To combat this, the dictionary was expanded by manually adding an extensive list of countries, nationalities and languages. Next, we performed automatic expansion with `word2vec`, using the best-performing model from Tulkens et al. (2016a). Finally, we

[8] An equivalent way (in this application) to capture this phenomenon would be to calculate log(types/tokens).

	Negative	Neutral
Skin color	✓	✓
Nationality	✓	✓
Religion	✓	✓
Migration	✓	✓
Country	✓	✓
Stereotypes	✓	
Culture	✓	
Crime	✓	
Race	✓	
Disease	✓	

Table 3: Overview of the categories in the racism-related dictionary

manually cleaned the word lists by removing irrelevant terms that were added automatically. In this stage of the research, the annotators had already seen the test data. However, no information from the test set was deliberately used in the expansion and filtering of the dictionaries. The presence of an unconscious bias, however, is possible. A downside of using dictionary-based models for the detection of racism is that they do not include a measure of context, and may therefore classify unrelated sentences as racist when they contain words that *might* be construed as racist, such as `black` or `sandpit`. The final dictionary consists of 3532 words divided into 10 categories. These categories are related to racist discourse and are subdivided into a negative and, optionally, a neutral subcategory. An overview of the categories can be found in Table 3.

To expand the recall of our dictionaries, two different kinds of wildcards were added to the beginning or end of certain words. The inclusive wildcard `*` matches the word with or without any affixes, whereas the exclusive wildcard `+` only matches words when an affix is attached (e.g. `moslim*` matches both `moslim` (Muslim) and `moslims` (Muslims), whereas `+moslim` will match `rotmoslim` (rotten Muslim) but not `moslim`). In our training corpus (which is skewed towards racism), the `+` will almost always represent a derogatory prefix, which is why it figures more prominently in the negative part of our dictionary.

In addition to our newly created dictionary, we also used Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001), which is a dictionary-based approach with both semantic categories (e.g. negative emotion words) and functional categories (e.g. pronouns). In the past, LIWC has been used to show correlations between authors' gender and personality and their writing style (Pennebaker 2011). Because some of LIWC's categories encompass certain features of racist language, e.g. use of pronouns and truth claims, we think it might be useful in trying to distinguish between various, more implicit, forms of racist discourse. Because our research concerns Dutch texts, we use the Dutch version of LIWC (Zijlstra et al. 2004).

3.4 Results

We will discuss different models' performance on the training data, obtained in tenfold cross-validation, and on the test data. We report precision, recall, and F-scores for the racist class, as this is the primary concern of the current research. To show the difference between performance on the racist class by itself and the global performance, we also report the macro-averaged F-scores for both classes.

In addition, we also report Area Under the Receiving Operator Characteristic Curve (ROC-AUC), which shows the probability of ranking a randomly chosen positive instance above a randomly chosen negative instance, thereby giving an indication of the overall performance of the models in classification.

	P	R	F	F Macro	ROC-AUC
All features	0.26	0.60	0.36	0.68	0.6
Sets					
Only Lexical	0.32	0.36	0.33	0.76	0.6
Only style-based	0.30	0.56	0.39	0.73	0.64
Character Trigrams	**0.51**	0.32	0.40	**0.82**	0.63
Dictionaries					
Racism-related	0.39	**0.70**	**0.50**	0.78	**0.73**
LIWC	0.26	0.69	0.37	0.66	0.64
Combined	0.38	0.68	0.49	0.78	0.72
Baselines					
Word Unigram	0.37	0.34	0.35	0.78	0.6

Table 4: Results on the training set.

We determined the optimal set of hyperparameters for our SVM based on an exhaustive search through the parameter space using tenfold cross-validation. Using this, we selected an *RBF* kernel, a C value of 1, a *gamma* value of 0, and scaled the C value of each class by the proportional class frequency.[9] We used the `scikit-learn` (Pedregosa et al. 2011) package throughout.

3.4.1 Results on the training data

We tested the models' performance on the training data using tenfold cross-validation. A detailed comparison of all scores can be found in Table 4. The baseline word unigram model obtains an F-score of 0.35 (std. dev. 0.07), with a slightly higher precision score.

The lexical model obtains an F-score of 0.33 (std. dev. 0.08), and does not outperform the word unigram baseline model, which shows that the features used, word trigrams, are probably not useful for classifying racist texts.

The model trained on stylistic features reaches a precision score of 0.30, a recall of 0.56 and F-score of 0.39 (std. dev. 0.07), thereby outperforming the unigram baseline. The observed pattern of performance confirms our expectations: the style-based model is not very precise, but has acceptable recall. The character trigram model shows an opposite pattern, achieving a high precision score (0.51), but lower recall (0.32).

The racism-related dictionary[10] model obtains a very high recall score of 0.70, and a lower precision score of 0.39. Its F-score of 0.50 (std. dev. 0.07) makes it the best-performing model on the racist class. The model trained on the LIWC features matches the racism-related dictionary model in recall, but is a lot less precise, achieving an F-score of 0.37 (std. dev. 0.07), which is quite a bit lower, but still outperforms the baseline. The combination of both the LIWC and racism-related dictionaries does not manage to outperform the racism-related model by itself, showing that the LIWC categories contribute negatively to the overall score. Nevertheless, the combined model might have more generalization potential, as the LIWC features are more general.

The model using all features, i.e. stylistic features, lexical features and dictionaries, achieved an F-score of 0.36 (std. dev. 0.06) for the racist class. This is surprisingly low, and indicates that there is most likely a high degree of competition between different sets of features.

[9] This corresponds to the `auto` setting in the `scikit-learn` implementation of the SVM.

[10] We note that, in our previous work (Tulkens et al. 2016b), this dictionary was called the 'Discourse dictionary'. However, to avoid ambiguity with other meanings of the word *Discourse*, we now refer to this dictionary and model as 'racism-related', as they are specifically related to *racist* discourse.

	P	R	F	F Macro	ROC-AUC
All features	0.24	0.02	0.04	0.62	0.50
Sets					
Only Lexical	0	0	0	0.62	0.50
Only style-based	0.37	0.50	0.43	0.65	0.59
Character Trigrams	**0.50**	0.27	0.35	0.70	0.59
Dictionaries					
Racism-related	0.46	0.49	0.47	**0.71**	0.64
LIWC	0.34	**0.63**	0.44	0.59	0.59
Combined	0.45	0.54	**0.49**	**0.71**	**0.65**
Baselines					
Word Unigram	0.45	0.20	0.28	0.68	0.56
Weighted Random Baseline	0.27	0.27	0.27	0.6	-

Table 5: Experiment 1: results on the test set.

3.4.2 Results on the test set

A detailed overview of the different models' performance on the test set can be found in Table 5.

On the test set, the model using all features reaches a ROC Area Under Curve (ROC-AUC) score of 0.50, and an F-score of 0.04 for the racist class. It nevertheless obtains an macro-averaged F-score of 0.62, showing that it is somehow unable to classify racist texts, but is able to classify non-racist texts to some degree.

Indeed, the model trained solely on word trigrams, i.e. the lexical model, does not manage to classify racist posts correctly. The fact that both the lexical model and the model trained on all features have a ROC-AUC score of 0.5 shows that they are not able to distinguish positive from negative instances in a reliable manner.

A reason for the bad performance of the lexical model could be sparsity. We have a relatively small number of instances, and hence word trigrams are not able to generalize well beyond the training set, since the word trigrams we encounter in the test set might not have occurred in the training set. In parallel, the trigrams account for a large number of features in the model trained on the combination of all features. Hence, the performance of this model is also degraded.

In stark contrast to the lexical model, the style-based model does show some degree of generalization, as the F-score on the racist class actually increases compared to cross validation setting, from 0.39 to 0.43, although the macro-averaged F-score decreases. The model using character trigrams as features also obtains reasonable performance, as its F-score of 0.35 is within one standard deviation of training performance.

The dictionary-based models all obtain scores which are within one standard deviation of the F-scores obtained on the training set, and thus are fairly robust. Comparable to the style-based model, the LIWC model obtains a better score on the test set than on the training set, although there is a drop in macro-averaged F-score. As predicted, the model which combines both LIWC and our own racism-related dictionaries shows a greater ability to generalize, and, as such, does not experience a drop in performance.

On the test set, all non-lexical models outperform the word unigram baseline in terms of recall and F-score of the minority class and AUC score.

4. Experiment 2

As the first experiment concerned the detection of racist comments on websites which were already skewed towards racism, we considered that two objections might be raised: first, the detection of racism on these pages might be too easy. In these venues, people might express racist opinions in rather explicit ways, as opposed to more implicit forms they might use in other contexts. Second, the nature of these pages might mislead our classification procedures, as we can expect most mentions of foreigners in these pages to be negative. Our classifier might, for example, learn that 'Muslim' by itself is a racist term because it never occurs in a non-racist context, simply because of the nature of the page itself. In order to alleviate these concerns, we collected a new, more neutral, set of comments, gathered from posts by Flemish newspapers on the same social medium. In this experiment, we use these comments as a new test set, to understand how well our classifier performs when transposed to a more neutral domain. The methodology of this second experiment is identical to the one from the first experiment (cf. Section 3.3).

4.1 Data Collection

To collect a more neutral test set, we turned towards the social media page of a well-known Belgian newspaper. From this page, we scraped 8 posts, all of which were links to news articles on the newspaper's website. The posts themselves only contained short utterances, usually paraphrases of the article to which the post linked. We retrieved all comments on these posts, just like we did for the first test set.

This resulted in a test set of 1138 comments - approximately 20% of the size of the training set. As the newspaper itself is neutral and does not exclusively post articles about subjects that could attract racist reactions, we argue that this set of comments will also be more nuanced and contain less instances of overt racism than the first test set. An important caveat is that, at the time of extraction, the arrest of a well-known terrorist figured prominently in the Belgian news. Because of this, many of the comments will still concern Muslim extremism, and possibly stereotypes regarding Muslims and extremism. Despite this, we argue that this set gives a more realistic view of racist reactions on general social media, thereby alleviating both of the problems above.

4.2 Annotation

The annotation of the new test set was performed by the same three annotators from Experiment 1, following the same guidelines as before. The same three possible labels were kept: a post could either be invalid, racist or non-racist.

25% of the test comments were annotated by all three annotators. The remaining 75% of the comments was equally divided among the annotators. The agreement was $\kappa = 0.63$ (pairwise average), which is higher than the score of $\kappa = 0.54$ in Experiment 1. This increase in agreement could be due to the fact that the annotators were more familiar with the task when performing it for the second time. The gold standard of the training and test corpus can be found in Table 6.

Concerning the pairwise comparison between the different annotators, it is interesting to note that annotator C is no longer the outlier, as was the case in Experiment 1, but has considerable overlap with annotator B ($\kappa = 0.7$). In the second experiment, annotator A was the outlier, having the lowest overlap with the other annotators, even with annotator B ($\kappa = 0.57$). This divergence could be explained by the fact that A no longer actively performed research on (the detection of) racist language, whereas the two other annotators continued discussing and studying the data. As in Experiment 1, A was still less inclined to annotate posts as racist when compared to both B and C. The following examples are some posts from the second test set which were annotated as racist:

7. Aan uwe [*sic*] schrijfstijl te zien ben jijzelf een Marokkaanse illegaal!!
   Going by how you write, you yourself are a Moroccan illegal immigrant!!

8. Ik persoonlijk moet daar ook niet van weten. Ik wil dat onze bevolking in stand blijft. Ik ben een trotse Belg.
   I personally don't want anything to do with this. I want our people to remain intact. I am a proud Belgian.

	# Test Comments
Non-racist	1019 (90%)
Racist	119 (10%)
Total	1138

Table 6: Gold standard corpus sizes Experiment 2.

9. Al meer dan 40 j zeer goede vriendin uit Tunesië,al tientallen keren op vakantie geweest...super familie,maar een relatie opbouwen met een moslim is quasi onmogelijk ,wij zijn TE verschillend.
I have been friends with someone from Tunisia for over 40 years, I have been on vacation there a lot of times... very nice family, but having a good relationship with a Muslim is almost impossible, we are simply TOO different.

10. Het zal wel weinig problemen zijn als ze van onder de zonnebank komen maar niet vanuit de woestijn. Maar ik denk niet dat het kleurtje de grootste tegenkanting zal zijn, maar het verschil in religie.
It doesn't matter to me when they use tanning beds, as long as they're not from the desert. I don't think the color is the biggest problem, but the difference in religion.

It is clear that these examples contain more implicit forms of racism when compared to the earlier batch of posts. There is almost a complete absence of direct slurs and name-calling, and when such terms occur ("Moroccan illegal immigrant"), they remain relatively descriptive, or non-derogatory. Also note that, in comparison to the previous set of comments, the authors seem to implicitly deny being racist. The author of Example 9 emphasizes the fact that she has been friends with someone from Tunisia, yet also mentions that it is impossible to bond with her family because they are Muslim (Bonilla-Silva 2002). Commenting on skin color, the author of Example 10 mentions that being non-white is not a problem, as long as it is from a tanning bed, and not 'from living in the desert'. Furthermore, skin color, although apparently an issue for the poster, is less of a problem than religion. Compared to the previous posts, there are less instances of conspicuous punctuation, such as sets of multiple exclamation or question marks. Also note the presence of tentativity in both Example 8 ("personally") and Example 10 ("I think"), which have been shown to be part of strategies for avoiding overt racist statements.

4.3 Results

As in the results section for Experiment 1 (Section 3.4), we will report the scores for the racist class (and for comparison, the macro-averaged F-score of both classes). As the results on the training data are the same as for the first experiment (cf. Section 3.4.1), we will only discuss the test runs of the second experiment.

4.3.1 Results

We tested the models from the first experiment on the new test set. The detailed results of all models can be found in Table 7.

The model using all features (lexical, stylistic, and dictionary-based) reaches a ROC-AUC score of 0.51, and an F-score of 0.07 for the racist class. The recall (0.04) is especially low. The macro-averaged F-score of 0.85, however, shows that the system based on all features is very good at classifying non-racist content as non-racist, but does not perform well in classifying racist content. This pattern of performance was also observed on the first test set, where the model only performed well on the majority class.

The results of the lexical model, which only uses word trigrams, were comparable to the results obtained on the first test set. The model obtains a very high score on the non-racist class, but does

	P	**R**	**F**	**F Macro**	**ROC-AUC**
All features	0.21	0.04	0.07	**0.85**	0.51
Sets					
Lexical	0	0	0	**0.85**	0.50
Only Style	0.14	0.44	0.21	0.72	0.56
Character Trigrams	0.22	0.20	0.21	0.84	0.56
Dictionaries					
Racism-related	**0.30**	0.63	**0.4**	0.65	**0.73**
LIWC	0.14	0.61	0.23	0.65	0.57
Combined	0.24	**0.67**	0.35	0.78	0.71
Baseline					
Word Unigram	0.15	0.10	0.12	0.83	0.52
Weighted Random Baseline	0.11	0.11	0.11	0.8	-

Table 7: Experiment 2: results on the test set.

not manage to classify racist content at all. The reason for the degraded performance of this model is the same as before: the lexical features are simply too sparse for a reliable transfer from train to test.

The style-based model obtains an F-score of 0.21, with low precision (0.14) and acceptable recall (0.44). This shows that solely relying on style-based features does not work well when transferring from one domain to another, even though the style-based classifier did perform relatively well on the first test set. Character trigrams do not transfer well to the new domain either, also experiencing a drop in performance, from 0.35 F-score to 0.21. It is worth noting that this model obtains a very high macro-averaged F-score of 0.84, rivaling the best performing models.

The dictionary-based models seem to be more robust: The racism-related model transfers relatively well to the new domain, obtaining a ROC-AUC score of 0.73 and an F-score of 0.40 for the racist class. Recall (0.63) is slightly lower than the score obtained by the combined system, but the precision score is higher (0.30). This suggests that grouping words and using a regular-expression-like mechanism increases generality, and that the racism-related dictionaries, which were handcrafted on the first domain, are more transferable than we initially thought.

The LIWC model does not transfer as well as expected when it comes to its performance for the racist class. This is probably due to a change in style. It is worth noting that the model obtains the same macro-averaged F-score of 0.65 as the racism-related model. As such, it is better at classifying non-racist content than the racism-related dictionary, which is expected.

Interestingly, as opposed to the results obtained in the first experiment, a combined model using both the general LIWC categories and the racism-related dictionaries does not outperform the model using only the racism-related dictionaries. The ROC-AUC score and F-score for the minority class both decrease slightly (to 0.71 and 0.35 resp.). Recall increases slightly compared to the racism-related model (0.67), whereas the precision score drops a little bit (0.24). Nevertheless, it obtains a higher macro-averaged F-score, which shows that it better accounts for non-racist content than the racism-related dictionary by itself.

In comparison to a basic word unigram model, all non-lexical models reach a higher F-score for the racist texts as well as a higher ROC-AUC score.

5. Discussion

Before discussing the performance of the different models and approaches, we note that an evaluation based on macro-averaged F-score, as is reported throughout the paper, might not be desirable in all setups. When using the detection tool in a moderator setup, we recommend evaluating based on recall (or weighted F-score), as the focus will then be on detecting all possible racist utterances, and manually filtering out false positives afterwards. In this paper, however, we have chosen to give a general idea (not yet based on a specific setup) of the detection systems' abilities.

First of all, the non-lexical models (i.e. all models except the one using all three kinds of features and the one using only word trigrams) outperform the word unigram baseline and weighted random baseline in terms of F-score for the racist class. This indicates that a more thorough analysis of racist discourse can lead to better results than the most basic approach, and that, although difficult, it is doable to automatically detect racist language in social media posts.

The dictionary-based approach, as presented in previous work (Tulkens et al. 2016b), appears to be the best approach to this specific classification task, and was more robust to the domain transfer, i.e. from the pages skewed towards racism to more general comments, than we initially expected.

On both test sets, the model using all three kinds of features (lexical, stylistic and dictionary-based) and the model using only lexical features (i.e. word trigrams) do not perform well. As their macro-averaged F-scores are clearly better than their scores for the racist class only, the models seem to only be able to classify non-racist content. Both models generalize better to the unseen data of the second test set, which makes sense, as the second test set contains more neutral utterances than the first set.

A system only based on stylistic features does not perform better than a dictionary-based approach either. While the style-based model performs quite well in the test phase of the first experiment, its performance significantly decreases on the second test set. A possible explanation for this difference in results between the two experimental setups is that in the first experiment, training and test data are very similar, as the comments are mined from the same social media pages and might involve the same set of authors, or authors with similar backgrounds (who are subscribers to the websites). In the second experiment, there is a discrepancy between training and test material, as the test set consists of reactions mined from a different web page. While we could argue that the stylistic differences between test and training set are minimal because both sets contain the same "kind" of utterances, i.e. social media comments, there will most likely *not* be a considerable overlap between the authors of the training and test material, which leads to a greater amount of stylistic differences.

In both experiments, the dictionary-based models perform best. The LIWC model performs worse, which may indicate that LIWC's categories are too general for this task. The model relying on dictionaries related to racist discourse holds best in the test phases of both experiments. A combined model, using both the specific (racism-related) categories and the general LIWC word lists, generalizes best to the unseen data of the first test set, but the transfer to the unseen data of the second test set was worse than expected. In the second experiment, however, it does outperform the racism-related model when it comes to macro-averaged F-score. The combined model therefore seems to be better at classifying neutral data than racist data, and neutral posts are more present in the second data set. Sparsity of racist comments in the second (more neutral and more realistic) test set could thus be a possible explanation of the decline in performance of the models which use LIWC in Experiment 2.

In terms of robustness, finally, the models using a dictionary-based approach perform best as well. Their performance on the test sets remains closest to their performance on the training data, often even within one standard deviation. These models therefore seem to generalize well to new data.

6. Conclusion

We presented a content-based and style-based approach to detecting racist discourse in Dutch social media. We performed two different experiments, which differ only in the social media pages from which the testing material was retrieved.

In both experiments, several classification models using the Support Vector Machine algorithm were trained on the same training set, consisting of Dutch comments retrieved from two public social media sites skewed towards racism, and annotated by three annotators. In the first experiment, the test set consisted of unseen comments retrieved from the same pages as the training data, whereas in the second experiment, a more neutral test set was collected, consisting of comments retrieved from the social media page of a newspaper.

Several models were tested in both experiments, using (different combinations of) three kinds of linguistic features: lexical, stylistic and dictionary-based features. A dictionary-based approach achieved the best and most robust results in this task. Different systems were tested, relying on different dictionaries and word categories, varying from general to specific for racist discourse. The most specific model, which is the best-performing model from Tulkens et al. (2016b), performed best on both test sets, reaching an F-score of 0.47 (exp. 1) and 0.40 (exp. 2) for the racist class and ROC Area Under Curve scores of 0.64 (exp. 1) and 0.73 (exp. 2).

In contrast to previous work in the same domain, we achieve high performance using hand-crafted features. This is in stark contrast to the work by Nobata et al. (2016), in which lexical features and character ngrams performed best, and dictionaries and style-based features underperformed. This can be attributed to the fact that they had access to a much larger dataset, comprising of some 2 million labeled comments from a Yahoo news forum.

Greevy and Smeaton (2004a) did not use a dictionary-based approach for classification, and instead used lexical features. Their corpus, however, consisted of longer texts, which, again, provide more data for the use of lexical features than our relatively small corpus could provide.

In future work we would like to experiment with techniques such as topic modeling for the semi-supervised classification of racist discourse. As Van Dijk (2002) indicates, the number of topics is greatly restricted when discussing foreigners in racist discourse. Hence, we would expect this bias to certain topics to show up in our dataset as well. Additionally, given enough data, topic modeling could allow for the detection of more implicit forms of racist discourse as well.

Concerning our dictionaries, we would also like to experiment with alternative ways of expanding them, as their current performance is promising. Topic modeling and related techniques could also provide ways of expanding these dictionaries. Another way to retrieve more words related to racist discourse is by tracking authors which we know (from our dataset) are inclined to post racist comments.[11]

7. Acknowledgments

We are very grateful towards Leona Erens and François Deleu from Unia for wanting to collaborate with us and for pointing us towards the necessary data. We thank the three anonymous reviewers for their helpful comments and advice.

References

Böke, Karin (1997), Die Invasion aus den "Armenhäusern Europas". Metaphern im Einwanderungsdiskurs, *Die Sprache des Migrationsdiskurses. Das Reden über "Ausländer" in Medien, Politik und Alltag, Opladen* pp. 164–193.

Bonilla-Silva, Eduardo (2002), The linguistics of color blind racism: How to talk nasty about blacks without sounding "racist", *Critical Sociology* **28** (1-2), pp. 41–64, SAGE Publications.

Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu (2012), Detecting offensive language in social media to protect adolescent online safety, *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)*, IEEE, pp. 71–80.

[11] We thank an anonymous reviewer for pointing this out.

Cohen, Jacob (1968), Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit., *Psychological bulletin* **70** (4), pp. 213, American Psychological Association.

Coupland, Nikolas (2010), "Other" representation, *Society and Language Use* **7**, pp. 241–260, John Benjamins Publishing.

De Smedt, Tom and Walter Daelemans (2012), Pattern for Python, *The Journal of Machine Learning Research* **13** (1), pp. 2063–2067, JMLR. org.

Discriminatie en Diversiteit: tijd voor een interfederaal actieplan tegen racisme (2013), *Technical report.*

Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati (2015), Hate speech detection with comment embeddings, *Proceedings of the 24th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 29–30.

Greevy, Edel and Alan F Smeaton (2004a), Classifying racist texts using a support vector machine, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 468–469.

Greevy, Edel and Alan F Smeaton (2004b), Text categorization of racist texts using a support vector machine, *7es Journées internationales d'analyse statistique des données textuelles.*

Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang (2016), Abusive language detection in online user content, *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 145–153.

Orrù, Paolo (2015), Racist discourse on social networks: A discourse analysis of Facebook posts in Italy, *Rhesis* **5** (1), pp. 113–133.

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

Pennebaker, James W (2011), *The Secret Life of Pronouns: What Our Words Say About Us*, Bloomsbury Publishing.

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001), Linguistic inquiry and word count: LIWC 2001, *Mahway: Lawrence Erlbaum Associates.*

Pollak, Senja, Roel Coesemans, Walter Daelemans, and Nada Lavrac (2011), Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining, *Pragmatics* **21** (4), pp. 674–683.

Quasthoff, Uta (1989), Social prejudice as a resource of power: Towards the functional ambivalence of stereotypes, *Wodak, R.(ed.), Language, Power and Ideology. Amsterdam: Benjamins* pp. 181–196.

Reisigl, Martin and Ruth Wodak (2005), *Discourse and discrimination: Rhetorics of racism and antisemitism*, Routledge.

Tulkens, Stéphan, Chris Emmery, and Walter Daelemans (2016a), Evaluating unsupervised Dutch word embeddings as a linguistic resource, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association (ELRA).

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans (2016b), A dictionary-based approach to racism detection in Dutch social media, *Proceedings of the LREC 2016 Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*, European Language Resources Association (ELRA).

Van Dijk, Teun A (1993), *Elite discourse and racism*, Vol. 6, Sage Publications.

Van Dijk, Teun A (2002), Discourse and racism, *The Blackwell companion to racial and ethnic studies* pp. 145–159.

Verschueren, Jef (1999), *Understanding pragmatics*, Oxford University Press.

Zijlstra, Hanna, Tanja Van Meerveld, Henriët Van Middendorp, James W Pennebaker, and Rinie Geenen (2004), De Nederlandse versie van de ‘linguistic inquiry and word count’(LIWC), *Gedrag & gezondheid* **32**, pp. 271–281.