The Early Modern Dutch Mediascape. Detecting Media Mentions in Chronicles Using Word Embeddings and CRF

Alie Lassche
Leiden University
a.w.lassche@hum.leidenuniv.nl

Roser Morante
VU University Amsterdam
r.morantevallejo@vu.nl

Abstract

While the production of information in the European early modern period is a well-researched topic, the question how people were engaging with the information explosion that occurred in early modern Europe, is still underexposed. This paper presents the annotations and experiments aimed at exploring whether we can automatically extract media related information (`source`, `perception`, and `receiver`) from a corpus of early modern Dutch chronicles in order to get insight in the mediascape of early modern middle class people from a historic perspective. In a number of classification experiments with Conditional Random Fields, three categories of features are tested: *(i)* raw and binary word embedding features, *(ii)* lexicon features, and *(iii)* character features. Overall, the classifier that uses raw embeddings performs slightly better. However, given that the best F-scores are around 0.60, we conclude that the machine learning approach needs to be combined with a close reading approach for the results to be useful to answer history research questions.

1 Introduction

It is well known from previous studies that early modern Europe witnessed an eruption of news, which was due to developments such as the burgeoning business of print, the professionalization of postal networks, and the invention of the newspaper (Pettegree, 2014; Rosenberg, 2003). While the production of information by various media is a well-researched topic, the question how people were engaging with this information explosion is still underexposed aside from a few studies (van Groesen, 2016; Blaak, 2009). Historians find it difficult to answer questions such as 'who owned book shelves?', 'who read which books?', and 'who bought newspapers?', since they have not yet found a suitable source and method to research these issues concerning the spread of information from the perspective of the consumer.

A source of information that might contain the key to answer these questions are early modern chronicles. A chronicle is a text from someone who keeps a record of events happening in their surroundings, who believes that these events are worth recording, and that the best way to structure this information is to do so chronologically (Pollmann, 2016). To fill their chronicle with useful information, early modern chroniclers used a wide variety of information sources, such as pamphlets, newspapers, official announcements, gossips, conversations, songs and other sounds, letters from friends or relatives, or the chronicler themselves functioned as eyewitness. All the sources that a modern chronicler had available constitute the early modern 'mediascape', a term coined by Appadurai (1990). Furthermore, the genre of the chronicle has remained rather stable over the centuries, and since many texts have been preserved in both public and private archives, a rather homogeneous collection of manuscripts still exists.

This is why chronicles are a suitable source for historians to get more insight in the reception of news and information. Two aspects are interesting to analyse from a history perspective: (i) the media that are mentioned by the author of the chronicle (such as 'my uncle', 'the newspaper', 'a rumour'), and *(ii)* the information that the chronicler is reporting on, which they have obtained from the media. These data are contained in the chronicles and traditionally close reading methods would be applied. However, in order to generalise across chroniclers, it is necessary to analyze as many chronicles as possible. Since close reading is a time consuming process, computational methods should be applied to automatically find the data.

This is the goal of this study. We focus on the first aspect (media) and present a computational approach to automatically extract media related information from Dutch early modern chronicles. This has been possible thanks to the availability of

Proceedings of LaTeCH-CLfL 2021, pages 1–10
Punta Cana, Dominican Republic (Online), November 11, 2021.

a corpus of 350 Dutch chronicles from the period 1500-1850 fully digitized. Our research question is: to which extent is it possible to automatically extract media mentions from a corpus of early modern Dutch chronicles? Addressing this issue is relevant because having this information available would allow historians to find out which books chroniclers owned, which newspapers they read, and in what other ways they received new information. It also provides insights in how the media use of chroniclers differs among each other. It would be possible to investigate, for example, how the mediascape of a chronicler from a large city such as Ghent, relates to that of a chronicler from a small provincial town such as Hoorn. Furthermore, it tells us something about how the preferences for certain media develop over time. The broader context of this study is to explore whether the use of computational methods can help historians to analyse the data concerning the mediascape of early modern chroniclers.[1]

Section 2 presents related work, Section 3 describes the data, Section 4 the methods, Section 5 the experiments, and Section 6 the results. Finally, in Section 7 we put forward some conclusions.

2 Related work

Historical text mining is an incipient field (Piotrowski, 2012). Regarding the Dutch language, historical text mining studies are limited. Stylometric methods have been applied in authorship verification tasks on medieval and early modern texts (Kestemont et al., 2017; Kestemont, 2018). Furthermore, Dutch early modern songs have been quantitatively researched (Lassche et al., 2019; Lassche, 2019). Corpora of historical Dutch newspapers have been subject to research in several studies, including Wevers and Verhoef (2018), Wevers et al. (2020), and Ros (2019). This work is the first to investigate chronicles on a large scale and with the use of text mining methods. Early modern Dutch chronicles have only been qualitatively researched up until now (Blaak, 2009; Pollmann, 2016, 2017). The same goes for chronicles in other languages, such as German, where only case studies on a handful chronicles have been published (Rau, 2002; Mauer, 2001).

Word embedding models (or word vector models) have recently become a popular way to describe a corpus of texts. Well known implementations are `word2vec` and `fasttext` (Bojanowski et al., 2017). Regarding the Dutch language, they have been used for relation evaluation and dialect identification (Tulkens et al., 2016), and to detect semantic change in Dutch newspapers (Wevers and Koolen, 2020; Wevers, 2019) and parliamentary debates (Lange, van and Futselaar, 2019). Word embeddings have been used as features in classification tasks (see for example Beelen et al. (2021)), but to the best of our knowledge, this has never been done for the historic Dutch language. Studies with corpora in other languages show that there are multiple ways to implement the word embeddings as features of a classifier. In Huang et al. (2015), every one hot encoding word representation is simply replaced with its corresponding 50-dimensional embedding vector. The authors of (Wu et al., 2015) compared three strategies for deriving distributed word representations in a corpus of English clinical texts. They base their method on earlier work (Guo et al., 2014), in which three different approaches for utilizing the embedding features are compared. In Wu et al. (2015) it is shown that the binarized word embedding features caused the largest improvement of the performance of the CRF system, while Guo et al. (2014) concluded that a distributional prototype approach performed the best. Bansal et al. (2014) perform agglomerative hierarchical clustering of the embedding vectors to take into account all dimensions of a vector simultaneously. Dařena and Süss (2020) use a neural network that is allowed to update the word vectors during training, since their Czech corpus from which a word embedding model was build, was rather small.

As far as we know, the current study is the first in which historic Dutch word embeddings are used as features of a classifier. We also do not know of any studies in other historical languages, in which a similar classification task is performed.

3 Data

In the context of the research project 'Chronicling Novelty. New knowledge in the Netherlands, 1500-1850', led by Judith Pollmann (Leiden University) and Erika Kuijpers (VU), a corpus of about 350 early modern Dutch chronicles has been built over

[1]The underlying code and materials of this paper are accessible via Github, see `https://github.com/awlassche/media-mentions-latech`.

the last three years.[2] The texts included in the corpus have the following characteristics: they are written between 1500 and 1850, they are organised chronologically, they cover events that happened in the lifetime of the author, and they focus on local events more than national, individual or familial.

About 100 of these chronicles had been published before as a contribution to a journal, on the initiative of an archive, or in the private domain, and were digitised by the DBNL (Digital Library for Dutch Literature). The other 250 chronicles were kept in libraries and archives throughout the Netherlands and Belgium. Every manuscript page was scanned, and afterwards transcribed with both the Handwritten Text Recognition tool Transkribus (Kahle et al., 2017), and the help of volunteers on the online crowd sourcing platform Vele Handen. Currently, the corpus contains about 70,000 pages.

In order to improve the searchability of the corpus of chronicles, a second group of volunteers is labeling useful information in the completely transcribed chronicles. They annotate both content and lay out in the chronicle pages, using labels including `date`, `location`, and `person name`, as well as `page number`, `lists and tables`, and `copied text`. Besides, a select group of volunteers has labeled media mentions in the chronicles. These are the annotations that are used in this study, which will be discussed in more detail in the next subsection.

3.1 Annotated media corpus

A group of four volunteers, all having an above average knowledge of the early modern Dutch language and culture, has been annotating media mentions in the chronicles. They were provided with extensive guidelines in which media mentions were explained.[3] To extract media related information, three types of annotations are relevant:

- a `receiver`, who is the person receiving information;
- a `source`, which is the instance bringing information to the receiver;
- a `perception`, which is the way in which the source is bringing information to the receiver.

[2]On `http://www.chroniclingnovelty.com/kronieken/`, an overview of the corpus can be found.

[3]The annotation guidelines are accessible via the Github repository.

The label `perception` has four possible attributes: `oral/heard`, `written/read`, `seen`, or `else`. See the following examples translated from Dutch into English:

1. 'This morning <source> mayor Vorsterman </source> came <perception: oral/heard> telling </perception> <receiver> us </receiver> that because of the disease, no one was allowed to be buried in the church.'
2. 'On 18 February <receiver> we </receiver> have <perception: written/read> seen </perception> in the <source> Amsterdammer Courant </source> of that day that Her Royal highness had given birth to a healthy and well-made Prince on the 16th at 11 in the evening in The Hague!'
3. '<receiver> I </receiver> have been to Broek and have <perception: seen> seen </perception> the Prince having dinner and walking through the village.'
4. '<receiver> They </receiver> <perception: oral/heard> heard </perception> how a farmer hanged himself.'
5. '<source> They </source> <perception: oral/heard> said </perception> that he had two letters in his pocket.
6. 'After this oddity had been <perception: seen> shown </perception> in Bruges for 8 days...'

The examples show that a `source` can be a person (1) or a printed text (2). In (5), the `source` is 'they', but this word also often functions as `receiver`, as can be seen in (4). The different perceptions are in most cases obvious ('telling' is `oral/heard` in (1), 'seen' is `seen` in (3)), but sometimes they are not. In (2), the word 'seen' suggests a `perception: seen`, but the context makes clear that this word should get the `perception: written/read`. Finally, the examples show that the mention of a medium does not necessary consist of the 3 labels: in (3) and (4) there is no `source`, in (5) there is no `receiver`, and in (6), there is only a `perception`.

Inter annotator agreement (IAA) was calculated at two moments during the process of improving the guidelines, using the balanced F-measure (Hripcsak, 2005) (see Table 1). After the first calculation of the IAA, the F-scores were analysed. They showed that the guidelines caused the most confusion among the annotators regarding the label `source`. Annotators found it hard to distinguish between the description of an event ('Our Aldermen Court were heard') and the mention of a medium ('We heard a strange rumour'). Guidelines were also not clear about self references of a

chronicler ('as I wrote on p. 23'). Some annotators interpreted this wrongly as a medium mention.

Table 1: IAA in the media annotation task.

	F-score phase 1		**F-score phase 2**	
	A1 - A2	*A2 - A1*	*A1 - A2*	*A2 - A1*
all	0.589	0.589	0.755	0.729
source	0.208	0.208	0.768	0.760
receiver	0.777	0.777	0.667	0.571
perception	0.707	0.707	0.754	0.699

The labels `receiver` and `perception` are less ambiguous: the receiver is often the chronicler referring to him- or herself, and a perception is usually a verb such as 'read', 'hear', 'tell', or 'see'. Annotators more often agreed on those. However, they also sometimes mixed up the `source` and the `receiver`, because the same word could have a different meaning in different contexts: in 'they heard', they is a `receiver`, but in a construction with 'they said', they is a `source`. In the updated guidelines, this definition of a `source` was more clearly explained, as well as the difference between a `source` and a `receiver`. Several examples were added, both in which media mentions appeared, as well as sentences which we considered not to contain media mentions. The F-scores obtained in the second inter annotator agreement after improvement of the guidelines made it clear that a lot of the confusion was cleared up: especially the F-score of the label `source` was much higher than before, as shown in Table 1.

In general, the F-scores show that we are dealing with a difficult annotation task. While annotating dates, person names and locations is something annotators hardly disagree on, labeling media mentions requires prior knowledge about the early modern Dutch society and language. The fact that the phrase 'de vliegende poste' (the flying post) is actually a reference to a carrier pigeon, and therefore should be labeled as a `source`, is something that is easily overlooked. Still, the final F-scores were considered good enough to let the volunteers make the annotations. About 25% of the total corpus of 350 Dutch chronicles is currently annotated with medium mentions, a percentage that is increasing everyday.

For the experiments reported in this study, we use four eighteenth century chronicles that are fully annotated with media mentions. Detailed corpus characteristics are presented in Table 2.

All experiments were performed on two different training and test sets. Set 1 contained the 12 volume chronicle from the late eighteenth century, written by Jozef Van Walleghem about his city Bruges (Van Walleghem, 2016). In set 2, three other eighteenth century chronicles were added (Abbing and S., 1794; Anoniem, 1795; Callion, 1789). For all experiments, of every chronicle, 70% was used for training, and 30% was used for testing.

Statistics about the datasets are displayed in Table 3. They show that set 2 is about 2.5 times bigger than set 1 regarding the number of tokens. This does not apply to the number of labels of `source-B`, `receiver-B`, `perception-B` and `perception-I`. This means that set 1 contains more labels per x tokens than set 2, in other words: the chronicle of Van Walleghem contains an above average number of media mentions. However, the labeled sources are longer in set 2: the number of `source-I` labels in set 2 is more than 4 times larger than in set 1.

4 Methods

The manually labeled data was used to train a CRF-classifier that is able to recognize media mentions in unseen chronicles. We use the Python library `sklearn-crfsuite` and the default parameters of CRF (Pedregosa et al., 2011). The classification task consists in labelling tokens with one of the following labels: `O` (no medium mention), `source-B`, `source-I`, `receiver-B`, `receiver-I`, `perception-B`, or `perception-I`.

Conditional Random Fields is a frequently used method in Named Entity Recognition (NER) labeling tasks, because CRF is able to deal with sequential data implicitly, without adding this information as features. Since the media mentions are tagged on token level using the IOB-tagging scheme, this task is also one in which labels of consecutive points can influence each other. There are, however, a few aspects in which this labeling task differs from a NER-task performed on modern texts, and complicates it.

Preprocessing of the data was limited: upper cases and punctuation were not removed, since these can be an indication of a media mention (consider the 'Gentsche Gazette'). Since chronicles were handwritten by authors from different areas, different time periods, and with different levels of literacy and dialects, there is a large variation in the vocabulary used. As a result of that, features such as POS-tags – which are often used in classi-

Table 2: Corpus characteristics.

set	chronicle	# pages	% labeled	# sources	# receivers	# perceptions
1 & 2	Brugge, Van Walleghem (1779-1800)	1165	17%	519	272	510
2	Gent, Callion (1780-1789)	248	100%	93	18	22
2	Hoorn, Abbing (1630-1794)	665	51%	241	94	53
2	Maastricht, Anonymous (1698-1742)	216	56%	63	74	36

Table 3: Statistics of tokens and labels per dataset.

set	1		2	
train/test	*train*	*test*	*train*	*test*
n tokens	107,351	46,006	284,160	121,781
source-B	373	122	772	251
source-I	486	152	2130	444
receiver-B	179	83	251	136
receiver-I	50	9	122	55
perception-B	335	134	470	235
perception-I	41	12	79	54

fication tasks that are performed on modern texts – can not be used in this task. Previous studies have shown, though, how the use of word embeddings as features of a classifier can improve its quality (Collobert et al., 2011; Wu et al., 2015; Bansal et al., 2014; Seok et al., 2016). Vectors of historical spelling variants are often each other's nearest neighbors. The hypothesis is therefore that using them as features will positively affect the quality of a model.

4.1 Features

We distinguish three categories of features: *(i)* word embedding features, *(ii)* character features, and *(iii)* lexicon features.

4.1.1 Word embedding features

We have used `fastText` to train a word embedding model (Bojanowski et al., 2017). In contrast to `word2vec`, `fastText` treats each word as composed of character n-grams, which enables the application to create a vector for a word that is not in the training data. The `skipgram` method was used, which was in previous studies shown to be better suited for this task than `cbow` (Ljubešić, 2018). The trained embeddings contained 100 dimensions. Three different word embedding models were used, in which the time span of the documents varied: model *a* was trained only on Van Walleghem's chronicle, model *b* was trained on chronicles that were finished between 1750 and 1800, and model *c* was trained on chronicles that were written between 1700 and 1850.

Regarding the obtained embeddings, two different ways of implementing them as features were tested: *(i)* RawEmb, and *(ii)* BinEmb. Both methods are also used by Wu et al. (2015). With RawEmb, the real values from the embedding matrix are directly used as feature weights for the classifier, without any post processing. The BinEmb refer to binarized embedding features, which are used to discretize the real-valued matrix and omit the insignificant dimensions. Given a word embedding matrix with the raw embeddings, the binarized embedding features are derived by converting the real-valued embedding matrix to another discrete-valued matrix with the discrete symbolic values in $[+, -, 0]$.

4.1.2 Character features

The character features that were calculated during training included the following features of a $token_n$, the previous $(token_{n-1})$, and the next $(token_{n+1})$:

- First 5 characters;
- Last 5 characters;
- Boolean value whether the token is a digit;
- Boolean value whether the token consists of only lower cases;
- Boolean value whether the token is a title.

4.1.3 Lexicon features

We also defined a boolean feature indicating whether the token occurred in a lexicon of media words. This lexicon consists of tokens that were extracted from the train corpus manually annotated data. Different lexicons were used in the two sets of experiments, consisting of `source` labeled words that appeared in the texts that were included in the specific set of training data.[4]

5 Experiments

We perform several experiments:

Experiment 1: Baseline. We performed three baseline experiments. In the first, only the lexicon

[4]The lexicons are available upon request.

Table 4: F-scores of all experiments.

Train/test	Model Embeddings	1. Baseline			2. RawEmb		3. BinEmb	
		lex.	*ch.*	*lex.+ch.*	*emb.+ch.*	*emb.+ch.+lex.*	*emb.+ch.*	*emb.+ch.+lex.*
set 1	Walleghem	0.573	0.569	0.575	**0.606**	0.600	0.581	0.571
	1750-1800				0.611	**0.612**	0.551	0.596
	1700-1850				0.594	**0.597**	0.586	0.593
set 2	1750-1800	0.346	0.359	0.359	0.391	**0.395**	0.367	0.367
	1700-1850				0.387	0.387	**0.402**	0.387

feature was used, in the second, only the character features were used, and in the third, both the lexicon feature and the character features were used.

Experiment 2: RawEmb. We performed two experiments with raw embeddings. In the first, only the character features and the word embedding features were used. In the second, the lexicon feature was added as well.

Experiment 3: BinEmb. We performed two experiments with the binarized embeddings. In the first, only the character features and the word embedding features were used. In the second, the lexicon feature was added as well.

6 Results

6.1 Performance of the classifiers

In Table 4, we present the F-scores of the classifiers for all experiments.[5] The F-scores were calculated with the function `metrics.flat_f1_score`. The label `O` was excluded from this score. The baseline experiments with solely the lexicon features show that using the lexicon with media mentions provides relatively high results.

Regarding the **baseline experiments** on train and test set 1, the best performance was achieved with a combination of the lexicon features and the character features. Adding the word embeddings as RawEmb features results in an increase of the F-score with around 4%. The differences between the scores of the experiments with and without the lexicon features are very small and negligible. Using the BinEmb version of the word embedding features, together with the lexicon features, results also in an increase of the F-score with around 4%.

The highest scores with train and test **set 1** were achieved with the RawEmb features obtained from the corpus of chronicles written between 1750 and 1800. Presumably, the word embedding model made from the chronicle of Van Walleghem solely contained not enough tokens to create a coherent model. The fact that the quality of the word embedding model of chronicles between 1700 and 1850 did not outperform the 1750-1800 model suggests that taking a too wide time span of texts enlarges the spelling variation, and does not improve the quality of a word embedding model. Over all, because of the relatively low number of tokens, decreasing the number of dimensions of the word embedding models to 75 or 50 might improve its performance.

Concerning train and test **set 2**, all experiments show that the system had a lower performance than the ones trained on set 1. However, adding word embedding features to the model causes a higher increase of the F-score than in set 1. This time, the highest scores with train and test set 2 were achieved with the BinEmb features obtained from the corpus of chronicles written between 1700 and 1850. The experiment with the lexicon features and the RawEmb features obtained from the corpus of chronicles written between 1750 and 1800 comes in second place.

Tables 5 and 6 contain the F-scores per label from the baseline experiments on set 2, with the lexicon features and the character features respectively. These numbers were obtained with the function `metrics.flat_classification_report`. They show that the classifiers obtain a better precision than recall.

Table 5: Precision and recall of experiment 1 (set 2, baseline, lexicon features)

	precision	recall	F-score	support
source-B	0.518	0.287	0.369	251
source-I	0.639	0.212	0.318	444
perception-B	0.707	0.298	0.419	235
perception-I	0.333	0.019	0.035	54
receiver-B	0.823	0.375	0.515	136
receiver-I	0.250	0.018	0.034	55

The two best performing experiments on set 2 are in bold in Table 4, and their detailed F-scores are displayed in Table 7 and 8. The precision and recall scores of the distinct labels show that the system obtains the highest precision in labeling

[5] We use an abbreviation for the used features. 'lex.' refers to the lexicon feature, 'ch.' to the character features, and 'emb.' to the embedding features.

Table 6: Precision and recall of experiment 1 (set 2, baseline, character features)

	precision	recall	F-score	support
source-B	0.493	0.263	0.343	251
source-I	0.616	0.245	0.351	444
perception-B	0.645	0.340	0.446	235
perception-I	0.167	0.019	0.033	54
receiver-B	0.803	0.390	0.525	136
receiver-I	0.200	0.018	0.033	55

`perception-B` and `receiver-B`. However, the low recall implies that the system is still missing a lot. It would therefore be useful to explore in future experiments the effect of different cut-offs for label probabilities, rather than using the default argmax for the various labels. Furthermore it would be worth to experiment in future research with other classifiers. A bidirectional long short-term memory system (BiLSTM) might for example be more suitable for this task. This model relies on neural networks and has shown in other studies to outperform the CRF-system when the level of non-standardness of the text increases.

In the current experiments, the very low scores for `perception-I` and `receiver-I` are probably caused by the scarcity of these labels. The scores for `source-B` and `source-I` show that both models gain a similar performance regarding this label, although different features were used. An evaluation at sequence level could provide more insight in possible tendencies.

Table 7: Precision and recall of experiment 2 (set 2, 1750-1800, RawEmb, lexicon)

	precision	recall	F-score	support
source-B	0.558	0.347	0.428	251
source-I	0.580	0.302	0.397	444
perception-B	0.643	0.345	0.449	235
perception-I	0.125	0.019	0.032	54
receiver-B	0.794	0.397	0.529	136
receiver-I	0.111	0.018	0.031	55

Table 8: Precision and recall of experiment 2 (set 2, 1700-1850, BinEmb)

	precision	recall	F-score	support
source-B	0.555	0.323	0.408	251
source-I	0.570	0.338	0.424	444
perception-B	0.647	0.366	0.467	235
perception-I	0.143	0.019	0.033	54
receiver-B	0.750	0.375	0.500	136
receiver-I	0.125	0.018	0.032	55

6.2 Error analysis

Manually comparing the gold labels with the predicted labels of the best performing systems shows that five types of issues occur frequently.

Errors in labeling sources with long spans. This became already clear from the numbers in Table 3, as we discussed earlier. The mention of a medium is rarely limited to one or two labeled words ('Gentsche Gazette', 'letters', or 'amendment'), but is often part of a phrase such as 'the amendment of our most honorable Lord Bishop' or 'letters from their Royal Highnesses the Governors General of the Netherlands'. The classifier is in most cases able to label words such as 'Gazette', 'letters' and 'amendment' as `source`, but does not label the surrounding words as `source`, while the annotators often do. When precision and recall is calculated on token level, this results in a low recall.

Labels `source-B` and `source-I` and the way in which annotators define the start of a medium mention. Some annotators have included articles or demonstrative pronouns in a medium mention, while others have not. This means that a word such as 'declaration' is in some cases labeled as `source-B`, and in other situations as `source-I`, while the pronoun 'this' is labeled as `source-B`. This again results in a mismatch, when calculating precision and recall on token level.

The third type of error concerns the so-called **lonely tokens**. When there is a sequence of labeled tokens, the model is better in predicting labels than when there is no labeled context. When a `source` such as 'decree' is mentioned without companion of a `perception` or `receiver` label, the model often does not label it as such. Using the lexicon feature does partially overcome this problem, but still there are infrequent tokens in the test set, which are not in the lexicon, and therefore not labeled as a `source`. The same issue occurs for a `perception` such as 'said', or a `receiver` such as 'they': when there are no labeled tokens in the close context, the tokens are often not labeled.

A fourth issue that also concerns the use of a lexicon, are the **ambiguous sources**. The term 'placard', or 'rumours', for example, is sometimes labeled as a `source`, but in other contexts, it is not interpreted as a `source` by the annotator. This inconsistency might confuse the classifier, thus improving the chance of a wrongly predicted label.

The final issue involves the fact that the **classifier labels in a more consistent way** than human annotators do. The word 'they', for example, refers in most cases to the `receiver` of a news item. However, the chronicler sometimes also uses it in other contexts, in which no media mention appears. Annotators did not label this word, but the model sometimes wrongly did. On the positive, from this uniformity of the model sometimes also follows that media mentions that were overlooked by the human annotator, are anyway labeled by the model. There are also a few cases in which the annotator confused a `receiver` with a `source`, but the computer assigned the correct labels.

6.3 Sequence level evaluation of the task

In Subsection 6.1 we presented the evaluation results at token level for each of the labels predicted. However, the full task involves finding the sequences of words that express sources, perceptions and receivers. Here we present results of evaluating at sequence level in terms of full match and partial match. In full match all the tokens in the sequence need to be correctly labeled, in partial match at least one token in the sequence needs to be correctly labelled. We evaluate the output of the two best performing classifiers on set 2 (based on the F-scores in Table 4).

Table 9: F-measures at sequence level of the two best systems for source (S), perception (P) and receiver (R), in terms of full match (fm) and partial match (pm). Column REL contains results on set2 1750-1800 RawEmb, emb+ch+lex; column BE contains results for set2 1700-1850 BinEmb (emb+ch).

	Precision		**Recall**		**F-score**	
	REL	BE	REL	BE	REL	BE
S fm	0.127	0.112	0.060	0.052	0.081	0.071
S pm	0.571	0.566	0.351	0.351	0.435	0.433
P fm	0.319	0.317	0.157	0.157	0.211	0.211
P pm	0.483	0.486	0.298	0.306	0.368	0.376
R fm	0.806	0.777	0.368	0.371	0.505	0.503
R pm	0.806	0.777	0.368	0.371	0.505	0.503

For `source` and `perception` the full match scores are much lower than the partial match scores, whereas this is not the case for `receiver`. This is due to the fact that `receiver` labels span over one token. The `receiver` obtains the highest scores, with a high precision, caused by the fact that there is less lexical variation in the expression of `receivers`. The `perception` obtains the lowest partial match scores, whereas the `source` gets the lowest full match scores. This is due to the fact that `source` labels tend to be longer sequences.

In the view of these results, with partial match F-scores not higher than 0.51 for `receiver`, 0.38 for `perception` and 0.44 for `source`, we can answer our research question: we can automatically extract from the chronicles some of the media related information, but the results are not high enough so as to perform the analysis of the chronicles fully automatically because too many media mentions are missed by the system. This is not surprising, given the amount of spelling and lexical variation present in the chronicles. The train corpus is not big enough for this type of machine learning experiments. Additional improvements need to be done to the system in order to improve its performance before historians can rely on it to obtain data to answer their research questions. We consider that the output of the current system can be used to perform exploratory studies. One of the possible improvements should focus on obtaining a higher recall, which now is much lower than precision. A higher recall would be desirable if the system would be used to select examples for the annotation of more data.

7 Conclusions

In this paper we have presented experiments aimed at extracting media related information from Early Modern Dutch chronicles with the purpose of investigating to which extent it is possible to perform this task automatically and to which extent are the results of automatic systems useful to facilitate history research. We have described the manual annotation task, which consisted in labeling sequences of tokens with the labels `source`, `perception` and `receiver`, considered by historians to be of relevance for their research. The IAA scores, with F-scores of around 0.73, show that the annotation task is well defined but that it is not easy. The label that produced more confusion was `source`, due to different interpretations of when to assign this label.

Four fully annotated chronicles were used to perform experiments with a CRF algorithm and different types of features that included several models of raw and binary word embeddings. We performed evaluations on token and sequence level. Both reveal that the systems obtain a much better precision than recall. In any case, the results of baseline experiments without word embeddings are

only 0.04 F-score lower than the experiments with word embeddings, indicating that the word embedding models are not rich enough, probably due to the size of the corpus used to generate them. Experiments with set 2, which contains more variety in authors, and therefore also in vocabulary, benefit more from word embedding features than experiments with set 1, where only one chronicle was used. The classifier with raw embeddings performs slightly better than the one with binary embeddings. As expected, the scores of the sequence level evaluation are lower than the scores of the token level evaluation.

In this paper we aimed at determining whether it was possible to automatically extract media related information from chronicles and whether the extracted information can be used by historians to answer their research questions. Since the sequence level partial match F-scores are not higher than 0.51 for receiver, 0.38 for perception and 0.44 for source, we conclude that the methods that we applied are not precise enough to extract the information and that it is not possible for historians to solely use such a system to obtain data to answer their research questions because too much information is missed. Both the data used to train the classifiers and the methods need to be improved. For future work we aim at experimenting with more algorithms and with larger corpora that include more chronicles, both for creating the embedding models and to train the systems.

Acknowledgements

Research for this paper was conducted at Leiden University and the Vrije Universiteit Amsterdam, and funded through the NWO VC project 'Chronicling Novelty. New knowledge in the Netherlands, 1500-1850'. We are thankful to the anonymous reviewers for their insightful comments, and to the annotators.

References

D.C.A. Abbing and H. S. 1794. Vervolg op de kroniek van Hoorn van D. Velius, eerste deel tot 1794.

Anoniem. 1795. Manuscripta wegens de stad Maestricht de anno 998 usque anno 1742.

Arjun Appadurai. 1990. Disjuncture and Difference in the Global Cultural Economy. *Theory, Culture & Society*, 7(2-3):295–310.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring Continuous Word Representations for Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, Maryland. Association for Computational Linguistics.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Jeroen Blaak. 2009. *Literacy in Everyday Life. Reading and Writing in Early Modern Dutch Diaries*. Egodocuments and History Series ; v. 2. Brill, Leiden ; Boston.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*.

Edouard Callion. 1789. Gentsche kronijke : 1525-1835 / door Edou. Callion. 1780-1789.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (almost) from Scratch. *arXiv:1103.0398 [cs]*.

František Dařena and Martin Süss. 2020. Quality of Word Vectors and its Impact on Named Entity Recognition in Czech. *European Journal of Business Science and Technology*, 6(2):154–169.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha, Qatar. Association for Computational Linguistics.

G. Hripcsak. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv:1508.01991 [cs]*.

P. Kahle, S. Colutto, H. Hackl, and H. Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Mike Kestemont. 2018. Stylometric Authorship Attribution for the Middle Dutch Mystical Tradition from Groenendaal. *Dutch Crossing*, 42(3):203–237.

Mike Kestemont, Els Stronks, Martine de Bruin, and Tim de Winkel. 2017. *Van Wie Is Het Wilhelmus? De Auteur van Het Nederlandse Volkslied Met de Computer Onderzocht*. Amsterdam University Press, Amsterdam.

Milan Lange, van and Ralf Futselaar. 2019. Debating evil: Using word embeddings to analyse parliamentary debates on war criminals in the netherlands. *Prispevki za novejšo zgodovino / Inštitut za zgodovino delavskega gibanja*, 59(1):140–156.

Alie Lassche. 2019. *Cultural Evolution in Dutch Early Modern Songs. Topical Fluctuations in the Dutch Song Database (1550-1750)*. Master's thesis, Utrecht University, Utrecht.

Alie Lassche, F.B. Karsdorp, and Els Stronks. 2019. *Repetition and Popularity in Early Modern Songs*.

Nikola Ljubešić. 2018. Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 156–163, Santa Fe, New Mexico. Association for Computational Linguistics.

Benedikt Mauer. 2001. *"Gemain Geschrey" und "teglich Reden": Georg Kölderer - ein Augsburger Chronist des konfessionellen Zeitalters*. Number 29 in Veröffentlichungen der Schwäbischen Forschungsgemeinschaft Reihe 1, Studien zur Geschichte des bayerischen Schwaben. Wißner, Augsburg.

F. Pedregosa, 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*, 12:2825–2830.

Andrew Pettegree. 2014. *The Invention of News: How the World Came to Know about Itself*. Yale University Press, New Haven ; London, England.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*, volume 5 of *Synthesis Digital Library of Engineering and Computer Science*. Morgan & Claypool, San Rafael, Calif.

Judith Pollmann. 2016. Archiving the Present and Chronicling for the Future in Early Modern Europe. *Past & Present*, 230(suppl 11):231–252.

Judith Pollmann. 2017. *Memory in Early Modern Europe, 1500-1800*, first edition edition. Oxford University Press, Oxford ; New York, NY.

Susanne Rau. 2002. *Geschichte Und Konfession: Städtische Geschichtsschreibung Und Erinnerungskultur Im Zeitalter von Reformation Und Konfessionalisierung in Bremen, Breslau, Hamburg Und Köln*. Number Bd. 9 in Hamburger Veröffentlichungen Zur Geschichte Mittel- Und Osteuropas. Dölling und Galitz, Hamburg.

R.S Ros. 2019. The birth of the foreign : A digital conceptual history of buitenland in dutch newspapers 1815-1914.

Daniel Rosenberg. 2003. Early Modern Information Overload. *Journal of the History of Ideas*, 64(1):1–9.

Miran Seok, Hye-Jeong Song, Chan-Young Park, Jong-Dae Kim, and Yu-seop Kim. 2016. Named Entity Recognition using Word Embedding as a Feature. *International Journal of Software Engineering and Its Applications*, 10(2):93–104.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. *arXiv:1607.00225 [cs]*.

Michiel van Groesen. 2016. Reading Newspapers in the Dutch Golden Age. *Media History*, 22(3-4):334–352.

Jozef Van Walleghem. 2016. *Merckenweerdigste voorvallen en Daegelijcksche gevallen Brugge 1779-1800*. Gemeentebestuur, Brugge.

Melvin Wevers. 2019. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. *arXiv:1907.08922 [cs, stat]*.

Melvin Wevers, Jianbo Gao, and Kristoffer Nielbo. 2020. Tracking the Consumption Junction: Temporal Dependencies between Articles and Advertisements in Dutch Newspapers. *Digital Humanities Quarterly*, 14(1).

Melvin Wevers and Marijn Koolen. 2020. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–18.

Melvin Wevers and Jesper Verhoef. 2018. Coca-cola: An icon of the american way of life. an iterative text mining workflow for analyzing advertisements in dutch twentieth-century newspapers. *Digital humanities quarterly*, 11(4).

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:1326–1333.