ONLINE APPENDIX
Linking historical ship records to a newspaper archive

Andrea Bravo Balado, Victor de Boer, and Guus Schreiber

Department of Computer Science,
VU University Amsterdam,
Amsterdam, the Netherlands
a.c.bravobalado@student.vu.nl,{v.de.boer, guus.schreiber}@vu.nl

This online appendix contains nine items. First, in Appendix A we show a sample of a common ship instance within the newspaper dataset. In Appendix B, we present the resulting confusion matrices from our text classification experiments and in Appendix C, a simple visualisation of the labeled instances. Furthermore, in Appendix D, E and F we present the surveys used for evaluation. Finally, a sample of the training set used and the stop words list used for text classification are shown in Appendix H and I, respectively.

A Ship instance example

We found that common ship mentions included the name of the ship along with the last name of the captain (either before or after), a port name and a date at the beginning of the sentence. An example can be found on Table 1.

Table 1. Example of an instance where the last name of the captain appears along with the ship name.

Ship Name	Grietina
Captain's last name	Sprik
Text title	BINNENGEKOMEN
Text	NEW-YORK, 6 Aug.; Albert, Meijer, Bremen. UITGEZEILD. KROON3TAD, 16 Aug.; Thorbecke, Witting, Kopcnh. - **Grietina, Sprik**, Bergen. NEW-YORK, 8 Aug.; zeilkl.: Ceres, Meuldijk, Antwerpen. RIO JANEIR'), 19 Julij; Maria, Lindquist, Abo. BUENUS-AYRES, 2 Julij; Industrie, Muller, Montevideo.

B Text classification confusion matrices

The confusion matrix for the Naive Bayes text classifier can be found in Table 2. Furthermore, the confusion matrix for the SMO text classifier can be found in Table 3.

Table 2. Confusion matrix for the Naive Bayes text classifier

Naive Bayes classifier			
		Predicted class	
		Negative (0)	**Positive (1)**
Actual Class	**Negative (0)**	24	15
	Positive (1)	0	11
Precision: 1		**Relevant documents:** 26	
Approximate recall: 0.42		**F1 Score:** 0.59	

Table 3. Confusion matrix for the SMO text classifier

SMO classifier			
		Predicted class	
		Negative (0)	**Positive (1)**
Actual Class	**Negative (0)**	17	18
	Positive (1)	0	15
Precision: 1		**Relevant documents:** 33	
Approximate recall: 0.45		**F1 Score:** 0.63	

C Proportion of labeled results by classifier

The portion of resulting labeled instances by the Naive Bayes classifier can be visualised in Figure 1. Additionally, the portion of resulting labeled instances by the SMO classifier can be visualised in Figure 2.

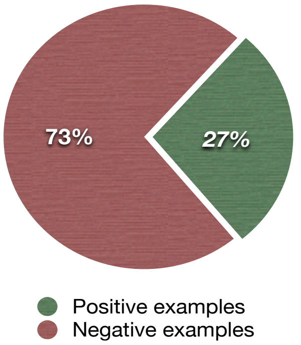


Fig. 1. Portion of labeled instances in the dataset by Naive Bayes text classification

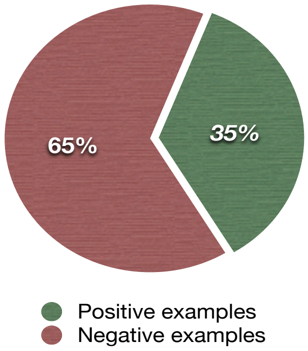


Fig. 2. Portion of labeled instances in the dataset by SMO text classification

D Article titles associated to positive and negative instances labeled by classifiers

After text classification, the article titles were analysed. The results can be visualised in Figures 3 and 4.

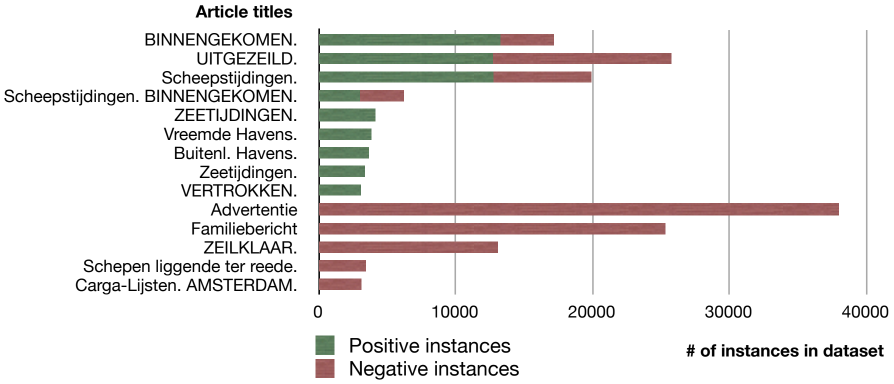


Fig. 3. Article titles associated to positive and negative text instances labeled by a Naive Bayes classifier

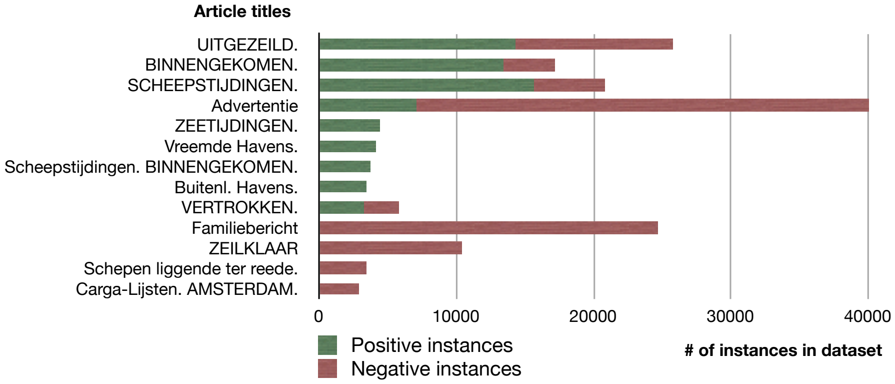
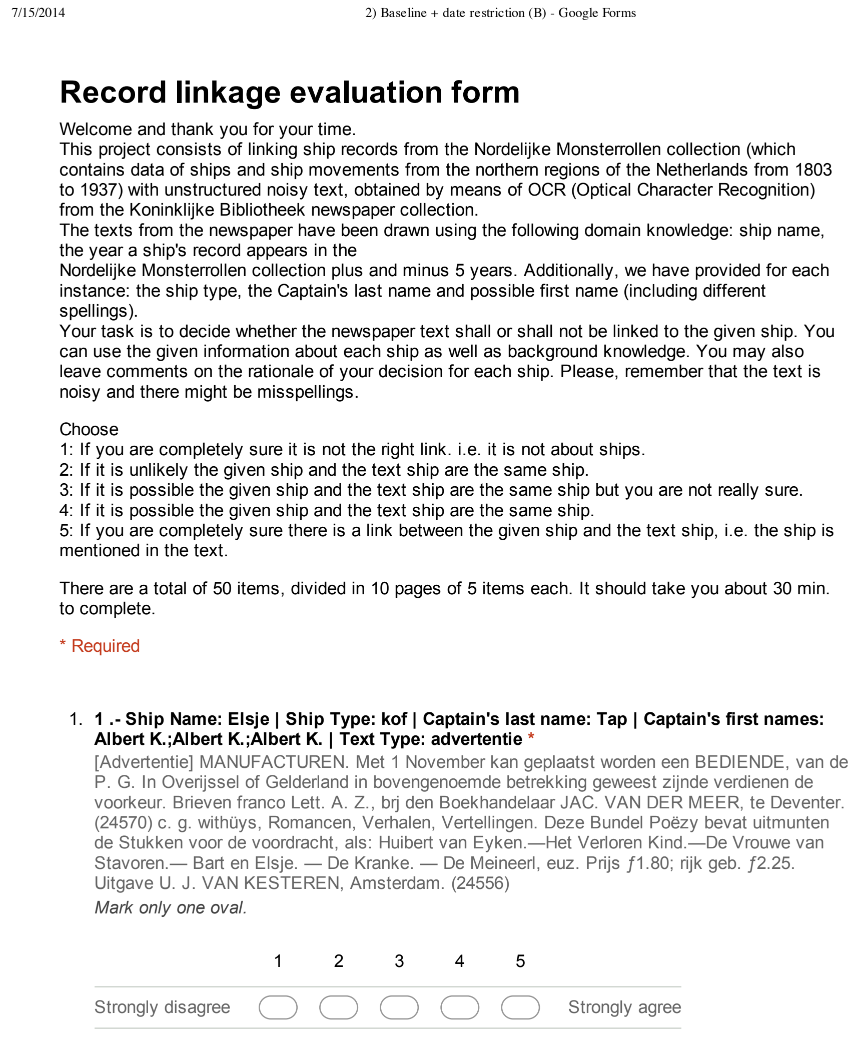


Fig. 4. Article titles associated to positive and negative text instances labeled by an SMO classifier

E Sample of the evaluation survey for baseline and Filters 1-2

The survey contained 50 items divided in 10 pages of 5 items each. It would take a rater around 30 minutes to complete. The surveys were managed using Google Drive Forms. The surveys for baseline, Filter 1a, Filter 1b and Filter 2 are available online.



7/15/2014 2) Baseline + date restriction (B) - Google Forms

Record linkage evaluation form

Welcome and thank you for your time.
This project consists of linking ship records from the Nordelijke Monsterrollen collection (which contains data of ships and ship movements from the northern regions of the Netherlands from 1803 to 1937) with unstructured noisy text, obtained by means of OCR (Optical Character Recognition) from the Koninklijke Bibliotheek newspaper collection.
The texts from the newspaper have been drawn using the following domain knowledge: ship name, the year a ship's record appears in the
Nordelijke Monsterrollen collection plus and minus 5 years. Additionally, we have provided for each instance: the ship type, the Captain's last name and possible first name (including different spellings).
Your task is to decide whether the newspaper text shall or shall not be linked to the given ship. You can use the given information about each ship as well as background knowledge. You may also leave comments on the rationale of your decision for each ship. Please, remember that the text is noisy and there might be misspellings.

Choose
1: If you are completely sure it is not the right link. i.e. it is not about ships.
2: If it is unlikely the given ship and the text ship are the same ship.
3: If it is possible the given ship and the text ship are the same ship but you are not really sure.
4: If it is possible the given ship and the text ship are the same ship.
5: If you are completely sure there is a link between the given ship and the text ship, i.e. the ship is mentioned in the text.

There are a total of 50 items, divided in 10 pages of 5 items each. It should take you about 30 min. to complete.

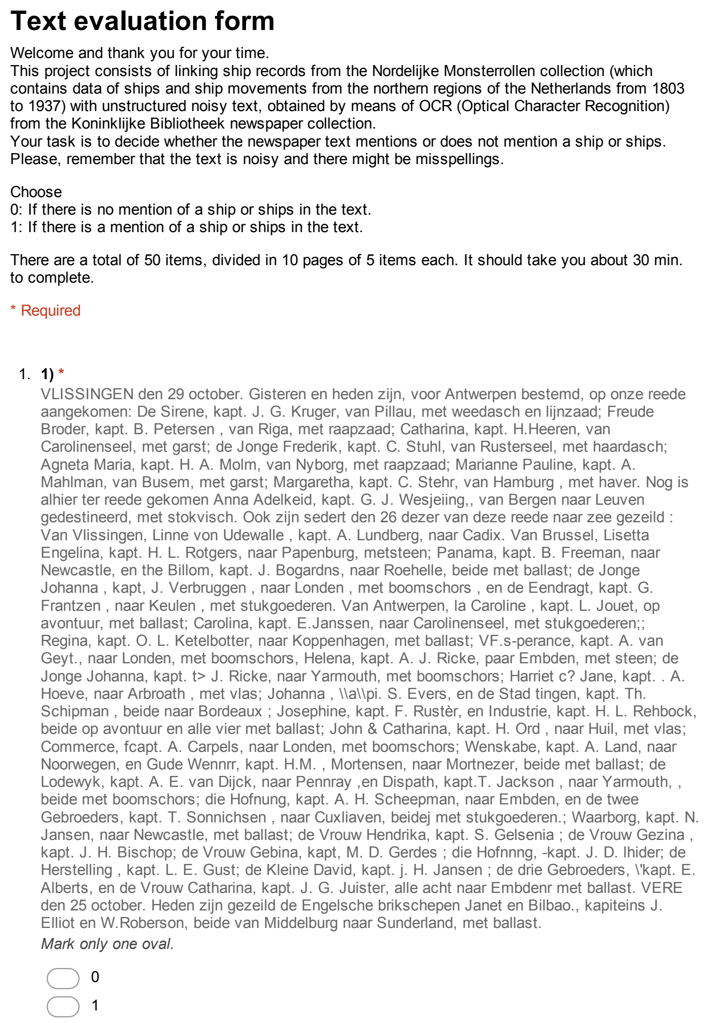
* Required

1. **1 .- Ship Name: Elsje | Ship Type: kof | Captain's last name: Tap | Captain's first names: Albert K.;Albert K.;Albert K. | Text Type: advertentie** *
   [Advertentie] MANUFACTUREN. Met 1 November kan geplaatst worden een BEDIENDE, van de P. G. In Overijssel of Gelderland in bovengenoemde betrekking geweest zijnde verdienen de voorkeur. Brieven franco Lett. A. Z., brj den Boekhandelaar JAC. VAN DER MEER, te Deventer. (24570) c. g. withüys, Romancen, Verhalen, Vertellingen. Deze Bundel Poëzy bevat uitmunten de Stukken voor de voordracht, als: Huibert van Eyken.—Het Verloren Kind.—De Vrouwe van Stavoren.— Bart en Elsje. — De Kranke. — De Meineerl, euz. Prijs ƒ1.80; rijk geb. ƒ2.25. Uitgave U. J. VAN KESTEREN, Amsterdam. (24556)
   Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

F Sample of the evaluation survey for text classification

Text classification experiments were manually evaluated by means of a survey which contained 50 items divided in 10 pages of 5 items each. It would take a rater around 30 minutes to complete. The surveys were managed using Google Drive Forms. The surveys for Naive Bayes and SMO are available online.



Text evaluation form

Welcome and thank you for your time.
This project consists of linking ship records from the Nordelijke Monsterrollen collection (which contains data of ships and ship movements from the northern regions of the Netherlands from 1803 to 1937) with unstructured noisy text, obtained by means of OCR (Optical Character Recognition) from the Koninklijke Bibliotheek newspaper collection.
Your task is to decide whether the newspaper text mentions or does not mention a ship or ships.
Please, remember that the text is noisy and there might be misspellings.

Choose
0: If there is no mention of a ship or ships in the text.
1: If there is a mention of a ship or ships in the text.

There are a total of 50 items, divided in 10 pages of 5 items each. It should take you about 30 min. to complete.

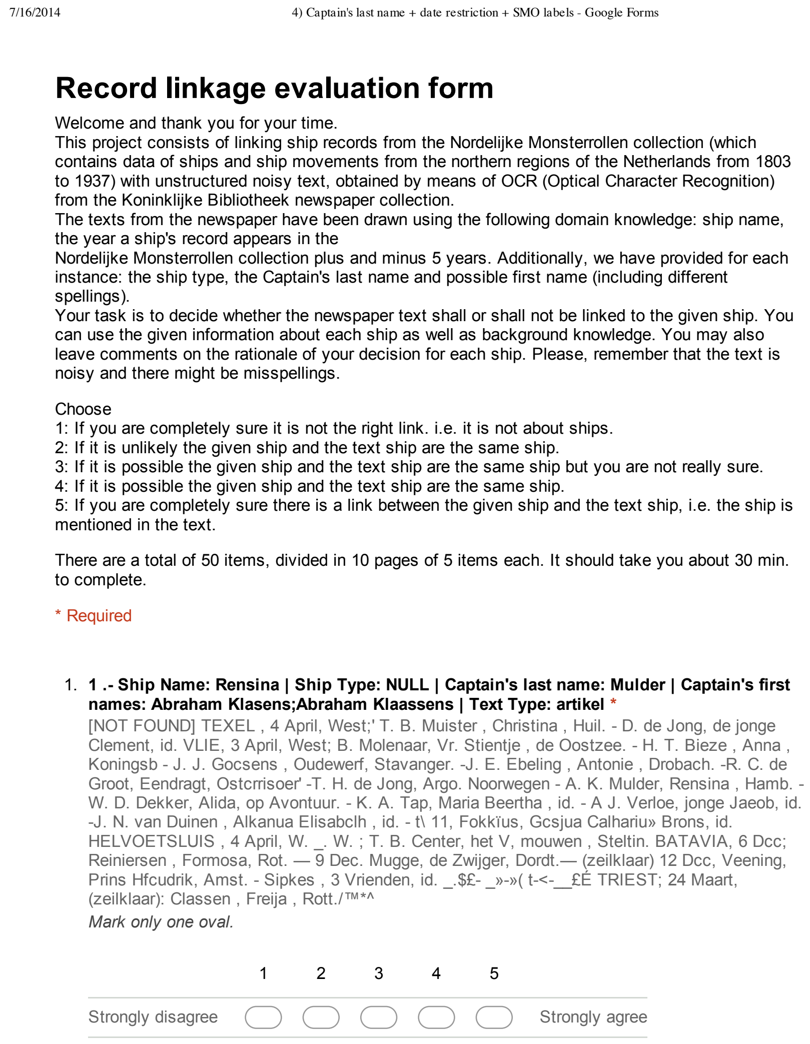
* Required

1. **1)** *
VLISSINGEN den 29 october. Gisteren en heden zijn, voor Antwerpen bestemd, op onze reede aangekomen: De Sirene, kapt. J. G. Kruger, van Pillau, met weedasch en lijnzaad; Freude Broder, kapt. B. Petersen , van Riga, met raapzaad; Catharina, kapt. H.Heeren, van Carolinenseel, met garst; de Jonge Frederik, kapt. C. Stuhl, van Rusterseel, met haardasch; Agneta Maria, kapt. H. A. Molm, van Nyborg, met raapzaad; Marianne Pauline, kapt. A. Mahlman, van Busem, met garst; Margaretha, kapt. C. Stehr, van Hamburg , met haver. Nog is alhier ter reede gekomen Anna Adelkeid, kapt. G. J. Wesjeiing,, van Bergen naar Leuven gedestineerd, met stokvisch. Ook zijn sedert den 26 dezer van deze reede naar zee gezeild : Van Vlissingen, Linne von Udewalle , kapt. A. Lundberg, naar Cadix. Van Brussel, Lisetta Engelina, kapt. H. L. Rotgers, naar Papenburg, metsteen; Panama, kapt. B. Freeman, naar Newcastle, en the Billom, kapt. J. Bogardns, naar Roehelle, beide met ballast; de Jonge Johanna , kapt, J. Verbruggen , naar Londen , met boomschors , en de Eendragt, kapt. G. Frantzen , naar Keulen , met stukgoederen. Van Antwerpen, la Caroline , kapt. L. Jouet, op avontuur, met ballast; Carolina, kapt. E.Janssen, naar Carolinenseel, met stukgoederen;; Regina, kapt. O. L. Ketelbotter, naar Koppenhagen, met ballast; VF.s-perance, kapt. A. van Geyt., naar Londen, met boomschors, Helena, kapt. A. J. Ricke, paar Embden, met steen; de Jonge Johanna, kapt. t> J. Ricke, naar Yarmouth, met boomschors; Harriet c? Jane, kapt. . A. Hoeve, naar Arbroath , met vlas; Johanna , \\a\\pi. S. Evers, en de Stad tingen, kapt. Th. Schipman , beide naar Bordeaux ; Josephine, kapt. F. Rustèr, en Industrie, kapt. H. L. Rehbock, beide op avontuur en alle vier met ballast; John & Catharina, kapt. H. Ord , naar Huil, met vlas; Commerce, fcapt. A. Carpels, naar Londen, met boomschors; Wenskabe, kapt. A. Land, naar Noorwegen, en Gude Wennrr, kapt. H.M. , Mortensen, naar Mortnezer, beide met ballast; de Lodewyk, kapt. A. E. van Dijck, naar Pennray ,en Dispath, kapt.T. Jackson , naar Yarmouth, , beide met boomschors; die Hofnung, kapt. A. H. Scheepman, naar Embden, en de twee Gebroeders, kapt. T. Sonnichsen , naar Cuxliaven, beidej met stukgoederen.; Waarborg, kapt. N. Jansen, naar Newcastle, met ballast; de Vrouw Hendrika, kapt. S. Gelsenia ; de Vrouw Gezina , kapt. J. H. Bischop; de Vrouw Gebina, kapt, M. D. Gerdes ; die Hofnnng, -kapt. J. D. Ihider; de Herstelling , kapt. L. E. Gust; de Kleine David, kapt. j. H. Jansen ; de drie Gebroeders, \'kapt. E. Alberts, en de Vrouw Catharina, kapt. J. G. Juister, alle acht naar Embdenr met ballast. VERE den 25 october. Heden zijn gezeild de Engelsche brikschepen Janet en Bilbao., kapiteins J. Elliot en W.Roberson, beide van Middelburg naar Sunderland, met ballast.
Mark only one oval.

○ 0
○ 1

G Sample of the evaluation survey for Filter 2 + Naive Bayes and Filter 2 + SMO

Filter 2 + Naive Bayes and Filter 2 + SMO were manually evaluated by means of a survey which contained 50 items divided in 10 pages of 5 items each. It would take a rater around 30 minutes to complete. The surveys were managed using Google Drive Forms. The surveys for Filter 2 + Naive Bayes and Filter 2 + SMO are available online.



7/16/2014 4) Captain's last name + date restriction + SMO labels - Google Forms

Record linkage evaluation form

Welcome and thank you for your time.
This project consists of linking ship records from the Nordelijke Monsterrollen collection (which contains data of ships and ship movements from the northern regions of the Netherlands from 1803 to 1937) with unstructured noisy text, obtained by means of OCR (Optical Character Recognition) from the Koninklijke Bibliotheek newspaper collection.
The texts from the newspaper have been drawn using the following domain knowledge: ship name, the year a ship's record appears in the
Nordelijke Monsterrollen collection plus and minus 5 years. Additionally, we have provided for each instance: the ship type, the Captain's last name and possible first name (including different spellings).
Your task is to decide whether the newspaper text shall or shall not be linked to the given ship. You can use the given information about each ship as well as background knowledge. You may also leave comments on the rationale of your decision for each ship. Please, remember that the text is noisy and there might be misspellings.

Choose
1: If you are completely sure it is not the right link. i.e. it is not about ships.
2: If it is unlikely the given ship and the text ship are the same ship.
3: If it is possible the given ship and the text ship are the same ship but you are not really sure.
4: If it is possible the given ship and the text ship are the same ship.
5: If you are completely sure there is a link between the given ship and the text ship, i.e. the ship is mentioned in the text.

There are a total of 50 items, divided in 10 pages of 5 items each. It should take you about 30 min. to complete.

* Required

1. **1 .- Ship Name: Rensina | Ship Type: NULL | Captain's last name: Mulder | Captain's first names: Abraham Klasens;Abraham Klaassens | Text Type: artikel** *
[NOT FOUND] TEXEL , 4 April, West;' T. B. Muister , Christina , Huil. - D. de Jong, de jonge Clement, id. VLIE, 3 April, West; B. Molenaar, Vr. Stientje , de Oostzee. - H. T. Bieze , Anna , Koningsb - J. J. Gocsens , Oudewerf, Stavanger. -J. E. Ebeling , Antonie , Drobach. -R. C. de Groot, Eendragt, Ostcrrisoer' -T. H. de Jong, Argo. Noorwegen - A. K. Mulder, Rensina , Hamb. - W. D. Dekker, Alida, op Avontuur. - K. A. Tap, Maria Beertha , id. - A J. Verloe, jonge Jaeob, id. -J. N. van Duinen , Alkanua Elisabclh , id. - t\ 11, Fokkïus, Gcsjua Calhariu» Brons, id. HELVOETSLUIS , 4 April, W. _. W. ; T. B. Center, het V, mouwen , Steltin. BATAVIA, 6 Dcc; Reiniersen , Formosa, Rot. — 9 Dec. Mugge, de Zwijger, Dordt.— (zeilklaar) 12 Dcc, Veening, Prins Hfcudrik, Amst. - Sipkes , 3 Vrienden, id. _.$£- _»-»(t-<-__£É TRIEST; 24 Maart, (zeilklaar): Classen , Freija , Rott./™*^
Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

H Sample of the training set for text classifiers in ARFF format

A labeled set was needed to train the classifiers for text classification. The labeled examples were obtained through the evaluation of the baseline as well as Filters 1-2. The training set contained 200 instances. A sample, featuring 4 instances of the training set can be found below.

```
@relation textEvaluation

@attribute evaluation {1,0}
@attribute text String

@data
0,"KONINGSBERGEN, H. J. Hazewinkel: Arendina Harmina: 2825
    schepels Graauwe Erwten, Order."
1,"TEXEL, HEDEN Vrijdag ochtend, 16 Sept.; binnengekomen: Aidina
     Anna Susanna, Schenk, Nickerie. - Anna en Arnoldina,van
    Wijk, Saramacca. LONDEN, HEDEN Vrydag 16 Sept. Consols op
    tijd 95} a}; Gren. Uitg. 8j a J; Buenos-Ayres 64 a 66;
    Spanje 1 pCts. 22* a } ; Dito j Certif. 5} a }. Overige
    onveranderd , maar vast. Dc beurs vertoont neiging tot ver.
    j dere rijzing. mm -f"
0,"HULL, Enchantress, T. Farr. Ñ 5525 stav. en 1897 boss. IJzer,
     E. S. de Jonge. HULL, Ocean Queen , C. Hardy. - 633 boss.
    IJzer, E. S. de Jonge. STOCKHOLM , Albertina, Lever. Ñ 2615
    staven en 96 boss. IJzer, E. S. de Jonge. STOCKHOLM, Cycloop
    , Takes. Ñ 485 stav. en 6 boss. IJzer, E. S. de Jonge."
1,"Amsterdam aang. 3 Nov. (IJkade) Danae, S, Patras. Stukgoed.
    Cargad. Hudig, Veder & Co. (Rietlanden) Dallon,B, Newcastle.
     Steenkolen. Cargad. Hudig, Veder & Co. 4 (Houthaven) Sigurd
    , S. Kroonstad. Hout, Cargad, Ruys & Co, (H.kade) Progress.
    S. Burryport. Steenkolen. Cargad. De Wed. Jan Salm & Meijer,
     Rynstroom, Ě, Huil. Steenkolen eu Stukgoed. Cargad. Holl.
    Stmbt.-Maatschij. 5 Koster, S. Newcastle. Steenkolen en
    Stukfo. d, Cargad. Hudig. ’eder & Co. IJ stroom, S. Londen.
    Stukgoed. Carg. Holl. Slmbt.-Maatsehij. Eapwing, S, Londen.
    t3tuks;oed. Carg. Nobel & Holtzapffel. (Houth.) Drammen, S.
    Christiania. Hout en IJzer. Cargad. B. J. van Hengel.
    Urnuiden 5 Nov. Wind Z.W.,hai _l: wind aangekomen 4 Progress
    , S. Burryport Sigurd, S. Petersburg Rijnstroom, S. . Huil
    Hermann, S. Archangel 5 Koster, S. Newcastle Drammen, S.
    Christiania IJstroom, S. Londen Lapwing, S, id. Hornfels, S.
     Riga Emden, S. Goole Orion, S. Hamburg vertrokken 4
    Waterland, S. Amble Milo, S. Bristol Alster, S, Hamburg Tem,
     S. Londen P/ulo, S. Malaga BŔta, S. West-Hartlepool De
    loodskotters zijn binnen; de stoomloodsboot doet loodsdien.
    t. Oostmalioru aang."
```

Listing 1.1. Sample of the training set in ARFF format

I Stopwords list used for text classification

For text classification, a stopword list was needed in order to ignore function and other unnecessary (for our purposes) words. Below is the list we have used for our experiments.

```
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, a, aan, acht, af, al, alle, alles,
     als, altijd, andere, april, augustus,
b, ben, bij,
c, d, daar, dan, dat, de, december, der, deze, die, dit, doch,
    doen, door, drie, dus,
e, een, eens, en, er,
f, februari,
g, ge, geen, geweest,
h, haar, had, heb, hebben, heeft, hem, het, hier, hij, hoe, hun,
i, iemand, iets, ik, in, is,
j, ja, januari, je, juni, juli,
k, kan, kon, kunnen,
l,
m, maar, maart, me, mei, meer, men, met, mij, mijn, moet,
n, na, naar, negen, niet, niets, nog, november, nu,
o, of, oktober, om, omdat, ons, ook, op, over,
p,
q,
r, reeds,
s, september,
t, te, tegen, tien, toch, toen, tot, twee,
u, uit, uw,
v, van, veel, vier, vijf, voor,
w, want, waren, was, wat, we, wel, werd, wezen, wie, wij, wil,
    worden,
x,
y,
z, zal, ze, zei, zes, zelf, zeven, zich, zij, zijn, zo, zonder,
    zou
```

Listing 1.2. Stopwords list