A data and analysis resource for an experiment in text mining a collection of micro-blogs on a political topic

William Black, Rob Procter, Steven Gray, Sophia Ananiadou

NaCTeM, School of Computer Science
University of Manchester
william.black@manchester.ac.uk
sophia.ananiadou@manchester.ac.uk

Manchester eResearch Centre (MeRC)
School of Social Sciences
University of Manchester
rob.procter@manchester.ac.uk

Centre for Advanced Spatial Analysis (CASA)
University College London
steven.gray@ucl.ac.uk

Abstract

The analysis of a corpus of micro-blogs on the topic of the 2011 UK referendum about the Alternative Vote has been undertaken as a joint activity by text miners and social scientists. To facilitate the collaboration, the corpus and its analysis is managed in a Web-accessible framework that allows users to upload their own textual data for analysis and to manage their own text annotation resources used for analysis. The framework also allows annotations to be searched, and the analysis to be re-run after amending the analysis resources. The corpus is also doubly human-annotated stating both whether each tweet is overall positive or negative in sentiment and whether it is for or against the proposition of the referendum.

Keywords: text analytics, social media, groupware

1. Introduction

The widespread adoption of new forms of communications and media presents both an opportunity and a challenge for social research (Savage and Burrows, 2007; Halfpenny and Procter, 2010). The rapid growth over the past ten years in the Web and the recent explosion of social media such as blogs and micro-blogs (e.g., Twitter), social networking sites (such as Facebook) and other 'born-digital data means that more data than ever before is now available. Where once the main problem for researchers was a scarcity of data, social researchers must now cope with its abundance. Realising the research value of these new kinds of data demands the development of more sophisticated analytical methods and tools. The use of text mining in social research is still at an early stage of development, but previous work in frame analysis and sentiment analysis indicates that this is an approach that has promise (Entman, 1993; Ananiadou et al., 2010; Somasundaran et al., 2007; Somasundaran and Wiebe, 2009; Wilson et al., 2009).

The project reported here is a case study of the use of text mining for the analyse of opinions manifest in twitter data. The key aim of the project is to explore the potential value to researchers of political behaviour of using text mining tools to extract the semantic content of twitter feeds, e.g. people, places, topics and opinions.

2. The AVtwitter Project

The AVtwitter project aims to provide social scientists with flexible text mining tools that they can use to explore social media content as primary data. A collection of 25K tweets was made over a 3-week period up to the recent UK referendum on the question of whether the Alternative Vote (AV) system should replace First Past the Post (FPTP) in elections to the UK parliament. For analysis, the corpus has been loaded in the Cafetière text analytics platform, which enables conventional text analysis (dictionary and rule-based named entity recognition, terminology discovery, sentiment analysis) to be carried out at the user's direction in a Web interface. Post analysis, the platform enables the user to search for semantic annotations by browsing.

3. The Corpus

The corpus comprises tweets sent in the period 10th April 2011 to the 7th May 2011 with a simple query '*AV*' as the selection criterion, harvested by SG. This seems to have worked quite satisfactorily as it has obtained greater coverage than would a restriction to topic-relevant hash tags such as #YestoAV. A very small proportion of noise exists, from one of two sources: Some tweets are in a language in which av is a preposition, and a slightly larger but still negligible proportion are using av as a 'text language' abbreviation for have, and are not on the topic of the alternative vote.

Table 1: Basic dimensions of the AVcorpus

Measure	Qty.
N. of tweets	24,856
N. of distinct followed sender IDs	18,190
N. of tweets referencing a @sender ID	7,698
N. of distinct @sender references with target in corpus	1,454

As Table 1shows, the corpus is of moderate size, and there are limitations due to the collection methodology. If we had wanted to focus exclusively on conversation structure as (Ritter et al., 2010), we would have filtered out those whose antecedents or followers are absent from the corpus. Nonetheless, we have the basis to analyze the structure of at least 1,454 distinct threads, as well as the corpus as a whole and the messages taken individually.

4. The Cafetière platform

The Cafetière platform was adopted for the AV twitter project, because being based on relational database corpus management, it is possible to conduct searches over

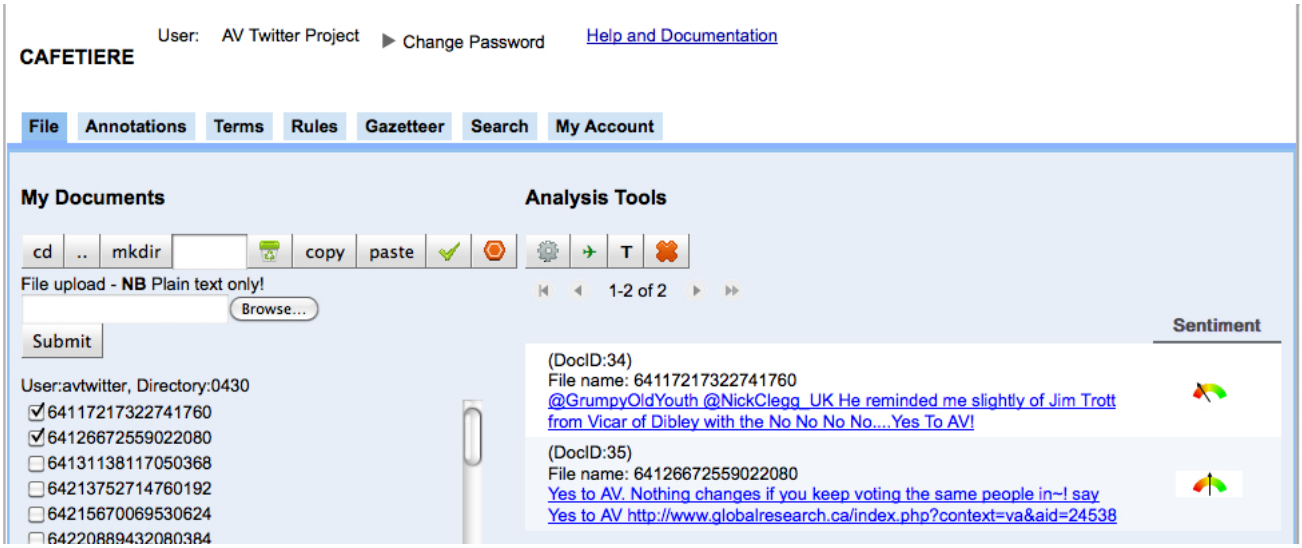

Figure 1: Cafetière analysis control panel showing links to individual document analysis and sentiment scores

the document metadata which comes with the twitter export, and metadata added in the course of text analytics applied to the textual content. The core of Cafetière is a UIMA-based (Ferrucci and Lally, 2004) suite of text analytic components, which cover basic text handling such as tokenization and sentence splitting, part of speech tagging, and then user-configurable analysis using dictionary matching and rule-based analysis. Earlier versions of the system are described in (Black et al., 1998; Vasilakopoulos et al., 2004). Based as it now is on UIMA, the components used for analysis are in principle interchangeable, but the user interface for 'self-service' text mining[1] does not currently allow the end user to change the low-level components or their configuration. Although deviance from normal orthography and spelling is a noted feature of twitter usage, it seems less of an issue with those joining the political debate, and we have used an un-adapted PoS tagger trained on a portion of the Penn Treebank corpus.

4.1. Corpus handling

A corpus of texts is held in the Cafetière system as a table in a relational database, the body text being held in a character large object field. Each user has their own private lightweight database created when they register. Users may manage their own corpora using the controls shown under the heading **My Documents** in Figure 1, which allow them to create and navigate between directories, and upload files for analysis. Files are handled according to their extension. Single .txt files are loaded into the currently open directory, and .zip files are unpacked after uploading to create a subdirectory within the currently open directory.

For the corpus of 24,856 tweets, prior to upload, we arranged the tweets into a directory for each distinct day in the period over which the data were collected, so as to avoid excess directory listing length. This is not currently a fully-automated procedure that users could replicate for themselves.

[1] Documentation and system are available at http://nactem3.mc.man.ac.uk/cafetiere

4.2. Analysis workflow

The main analysis workflow comprises a UIMA pipeline of processes:

1. Sentence splitting
2. Tokenization
3. PoS tagging
4. GeoNames lookup of place names
5. Dictionary lookup
6. Rule-based phrasal analysis

The sentiment lexicon is applied during the dictionary lookup stage, and sentiment-bearing words and phrases are just one category of many that can be looked up at a time.

Tokenization Tokenization has been amended to cater for the twitter corpus. Tags of the form @follower and #hash as well as URLs are treated as single tokens.

This may not be the last word on the matter, since we now consider it interesting to analyze @follower and #hash tags into component parts, since these tags often have real word boundaries indicated with 'CamelCase'. The parts of such a tag often contain sentiment-bearing words which are currently not exposed to dictionary lookup. For an example, see the tag '**@GrumpyOldYouth**' that appears in the first tweet that is visible in Figure 1.

PoS Tagging The part of speech tagger used in the pipeline is JTBL, an implementation of Brill's transformation-based rule-learning algorithm, which is available from Sourceforge. This tagger uses human-readable rules, a dictionary and a part of speech guesser based on suffix patterns. All of these resources can be modified to compensate for observed failures to deal with a particular corpus without retraining.

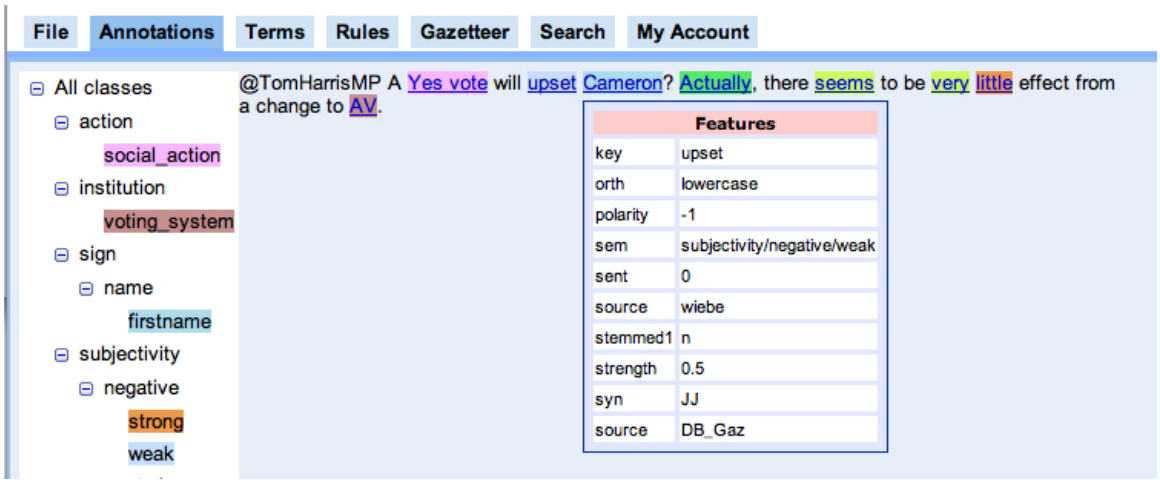

Figure 2: Annotation browser showing annotation popup and key

GeoNames lookup Place references feature in the corpus, and we have an established Cafetière annotator for GeoNames. Geographical names overlap to a great extent with names for people and other entities, and some disambiguation is needed. One heuristic we have in place is that we exclude all non-UK place names from consideration, but that is only reasonable because of the scope of the topic defining the current corpus.

Dictionary lookup Whilst many text mining pipelines use multiple dictionaries where each is a list of items in a single category, Cafetière uses a single relational database table[2] to store entries in all categories, each of which has a semantic class assignment, a preferred form (in the case of proper names or domain terms), and optionally any number of feature names and values. Figure 2 shows a detail view of an annotation that has been created by dictionary lookup. A textual format for dictionary entries allows lists of items to be assembled and uploaded in bulk, and there is a gazetteer editor accessible from the eponymous tabbed pane.

For the AV twitter corpus, the dictionary (also known as a *gazetteer* in the system documentation) contains an extensive lexicon of subjective expressions and smaller numbers of terms of specific interest in the domain of British electoral politics.

The GeoNames component uses the same dictionary technology, but as its content comes from a single source, it has been encapsulated as a separate UIMA annotator, which we run before the domain-specific dictionary annotator.

Rule-based analysis Cafetière supports phrasal analysis beyond the dictionary by means of a rule-based annotator. Production rules successfully applied create phrases of one or more text units which can be either tokens or phrases previously created by either a dictionary annotator or previous rules.

These rules describe both phrases and their constituents as sets of feature-value pairs with Prolog-style unification of variables. The rules may be context-sensitive, in requiring the presence of constituents before or after a phrase, but which do not form part of it. The rule formalism is explained, with examples, in the system documentation.

Rules are applied in sequence, so that the analysis is deterministic. The formalism is therefore more suited to syntactic analysis up to the level of chunking, or to named entity recognition, than to full parsing.

In the analysis of the AV corpus, rules are used to contextually constrain the applicability of the items from the sentiment lexicon, including reversing polarity scores based on contextual items.

Context-sensitive sentiment analysis can be achieved by rules that promote or demote the sentiment scores of looked-up words or phrases, or by the creation of phrases from parts that are not sentiment-bearing out of context.

Post-processing Sentiment scoring is undertaken after the output of the UIMA analysis has been written to searchable database tables, and scores are computed by aggregate SQL queries.

It is simply for convenience that we currently compute sentiment scores outside of the UIMA pipeline, but there are other types of analysis for which the UIMA framework is not suitable. When we mine the corpus for topical phrases (See Section 7.1.), this analysis is carried out on the corpus as a whole, not independently on individual texts. The UIMA common analysis structure (CAS) that is created as a result of the pipeline's analysis steps applies to a document at a time and is destroyed when the next text is input. Hence, any corpus-level analysis must be completed outside of the CAS.

5. User-configurable analysis

In the Cafetière Web interface, the user may upload and edit text for analysis, and resources with which to analyze those texts, in their private space on a server. Text files are uploaded to a single http file upload destination, and the system disposes of the files according to their file extension. Files of extension .txt are treated as data files, and they

[2] and a related prefix table to facilitate lookup of multi-word tokens

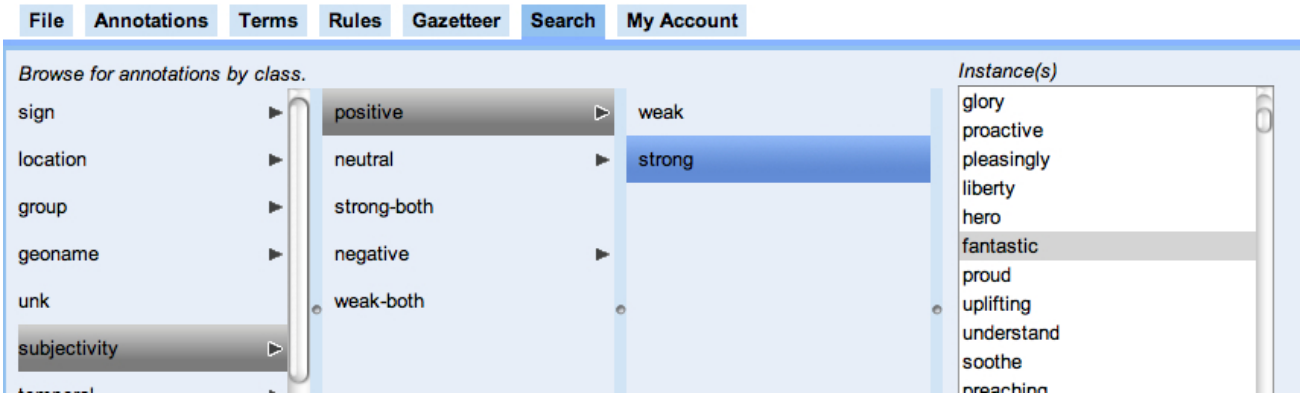

Figure 3: Annotation search by class and instance browsing

are placed in whichever 'directory' is currently open. Files of extension .gaz are treated as gazetteer files, and become part of the dictionary used for named entity-style analysis. The format of .gaz files is outlined in the on-line system documentation. In addition to uploading already compiled gazetteer files, the system allows the user to add and amend individual entries. Files of type .rul are context-sensitive syntactic/semantic rules that allow the creation of annotations on the basis of the satisfaction of feature constraints on their daughters, and if desired, on contextually adjacent text units. Twitter data is obtainable not as single files per text, but in the form of CSV files, in which the text column is complemented with metadata including the date, sender, sender profile, geo tag, etc. Web Cafeteire does not currently provide a facility to automatically upload such a file and split it into individual messages, but a batch update was conducted. For ease of reference in the interface, each distinct send date was placed in a directory of its own. For the analysis of the AV twitter corpus, we have concentrated initially on sentiment analysis, based initially on the MPQA sentiment lexicon (Wilson et al., 2009). The sentiment lexicon has been converted to the Cafetière .gaz format and this has been augmented with rules to take some account of context.

6. Corpus Annotation

In order to explore sentiment analysis in the corpus, each of the tweets has been annotated by two social science graduate students, who assigned each tweet two labels: one whether it expresses positive, neutral or negative sentiment towards the topic of the message, and secondly whether the writer was expressing an opinion for or against the proposition of the referendum. The agreement between the annotators (8 in total, working in pairs) has been computed at 82.43% for the for/against decision, but for the sentiment labelling, exact agreement stood at 49.15%, and agreement to within one point on the Likert scale, at 84.36%.

7. System Annotation

Sentiment annotation by the system has been computed with two alternative baseline conditions: one in which direct lexical matches only are used, and one in which various contextual factors are taken into account. In the first condition, the system produces a very different distribution to the human annotators, with over 70% positive sentiment, 1% negative, and the balance neutral. This is considered anomalous, as the topic of a referendum includes discussion of the proposition of voting Yes or No, both of which occur in the MPQA lexicon, and quoting such expressions does not imply expressing them subjectively. In the second condition, expressions involving Yes and No are excluded from the respective sentiment scores, as are a small number of words which have an auxiliary verbal sense that is not sentiment expressive (e.g. hope, might) and a nominal sense that is evaluative. This condition gave rise to a drastically different distribution of positive and negative sentiment (24% and 5% respectively, with the balance neutral). The prediction of sentiment scores and indeed of the for/against AV orientation of the tweets remains as work to be done. The methodology will be to use the human-annotated corpus for training with a hold-out set retained for testing. As both of the baseline results have given a strong balance of positive over negative scores, we will initially focus in the training set on the subset where human annotators have assigned a negative score and the system has not. This activity is under-way, and we have currently started to identified a range of expressions that are considered to hold negative connotations in the political sphere, when they are more neutral in other contexts. There are also cultural differences between the US and Britain in the subjective loading that different expressions bear, and the MPQA lexicon was developed in an American context.

7.1. Unsupervised Topic Analysis using TerMine

The UIMA-based text-mining pipeline is designed to carry out a document-by- document analysis of each text in a corpus. In a corpus such as the AV twitter collection, it is also of interest to be able to capture an indication of the semantic content of the corpus at a collection level. One tool at our disposal for this purpose is TerMine (Franzi et al., 2000), an implementation of which has been incorporated in the Web Cafetière toolset. A UIMA pipeline up to part of speech tagging is run as a preprocessor to TerMine, which then computes its C-value statistic on the distribution of terms from the whole corpus, including those that overlap. Ta-

Table 2: Top multi-word terms in three categories, as discovered by TerMine

Rank	AV slogans	C-value	Rank	People	C-Value	Rank	Other topics	C-Value
1	vote yes	888.94	9	david cameron	226.74	8	second choice	247.24
2	av campaign	541.07	11	nick clegg	192.83	13	lib dem	169.09
3	av referendum	497.36	16	nick griffin	133.11	14	second preference	155.96
4	av vote	336.48	24	eddie izzard	99.57	18	polling station	126.01
5	voting yes	321.59	29	ed milliband	81.05	19	fairer votes	117.10

ble 2 shows the top 5 multi-word terms as discovered by TerMine within the AV corpus in each of three categories.

8. Search Facilities

To support the social science users of the corpus, search facilities are provided (Figure 3 illustrates) where the annotations can be browsed for by category, and then by instance, leading to a search results list, where the annotations of the analysis (named entity and sentiment) can be viewed in a highlight viewer with feature popups.

9. System Availability

The system is currently accessible at http://www.nactem.ac.uk/cafetiere. To view the analyzed AV data, log in as the user avtwitter with the password yn2av. For up-to-date news about analysis resources for the AV corpus, follow the link to Help and Documentation, and look for the heading Social Media Analysis.

10. Further Work

We made reference above to the text analytic development and evaluation that remains to be done. Also planned are various minor augmentations to the Web- based analysis environment that have suggested themselves in the course of working with the twitter data. These include the facility to import one's own corpus of twitter data in CSV format, and the facility to exploit the output of TerMine in the creation of dictionary entries for NER.

10.1. Twitter in Argo

To ameliorate the problem that Cafetière supports only a single, albeit user-customisable, workflow, we plan to port the corpus and its existing analysis resources to the Argo platform (Rak et al., In Press; Rak et al., 2012) in the near future. This will allow for users easily to experiment with alternative modules for tokenization and tagging, as well as the dictionary and rule-based components that can be amended by users of Cafetiere. Since Argo also provides annotation editing and the training of CRF models, a range of different analysis approaches will be possible.
Also planned is a corpus reader component that will allow users to make their own collections from live twitter feeds on topics of their own choosing.

11. Conclusion

A corpus of just under 25,000 tweets on a single political topic (the referendum held in 2011 to determine whether Britain should adopt the Alternative Vote for parliamentary elections). This corpus is managed in the Cafetière Web-based system for text mining, and demonstration linguistic resources have been created for sentiment analysis and named entity analysis. The topics and key phrases used by those tweeting about the topic can be explored using TerMine, and the search facilities allow for the selective location of annotations based on their semantic class.

12. Acknowledgements

The software development and text analysis was funded by the JISC's grant to NaCTeM. The human annotation of the corpus was funded by methods@manchester's grant to MeRC, and the annotation itself carried out by Rosalynd Southern, Stephanie Renaldi, Jessica Symons, Paul Hepburn, Stephen Cooke, Jasmine Folz, Jan Lorenz, Stephanie Doebler, Patricia Scalco and Jinrui Pan.

13. References

Sophia Ananiadou, Paul Thompson, James Thomas, Tingtin Mu, Sandy Oliver, Mark Richardson, Yutaka Sasaki, Davy Weissenbacher, and John McNaught. 2010. Supporting the education evidence portal via text mining. *Philosophical Transactions of the Royal Society A*, 368(1925):3829–3844, August.

William J. Black, Fabio Rinaldi, and David Mowatt. 1998. Facile: Description of the NE system used for MUC7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*, Fairfax, VA, May.

R.M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

K. Franzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2):117–132.

P. Halfpenny and R. Procter. 2010. The e-Social Science research agenda. *Philosophical Transactions of the Royal Society A*, 368(1925):3761–78, August.

Rafal Rak, Andrew Rowley, and Sophia Ananiadou. 2012. Collaborative Development and Evaluation of Text-processing Workflows in a UIMA-supported Web-based Workbench. In *Proceedings of LREC 2012*, Istanbul, May.

R. Rak, A. Rowley, W.J. Black, and S. Ananiadou. In Press. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: The Journal of Biological Databases and Curation*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Savage and R. Burrows. 2007. The Coming Crisis of Empirical Sociology. *Sociology*, 41:885–899.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

Swapna Somasundaran, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. 2007. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in Online Discussions and the News. In *International Conference on Weblogs and Social Media (ICWSM-2007)*, Boulder, Colorado, March.

Argyris Vasilakopoulos, Michele Bersani, , and William J. Black. 2004. A suite of tools for marking up textual data for temporal text mining scenarios. In *LREC 2004*, Lisbon, May.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, September.