Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study

Eman M.G. Younis
Faculty of Computer and Information
Minia University, Egypt

ABSTRACT
Recently, Social media has arisen not only as a personal communication media, but also, as a media to communicate opinions about products and services or even political and general events among its users. Due to its widespread and popularity, there is a massive amount of user reviews or opinions produced and shared daily. Twitter is one of the most widely used social media micro blogging sites. Mining user opinions from social media data is not a straight forward task; it can be accomplished in different ways. In this work, an open source approach is presented, throughout which, twitter Microblogs data has been collected, pre-processed, analyzed and visualized using open source tools to perform text mining and sentiment analysis for analyzing user contributed online reviews about two giant retail stores in the UK namely Tesco and Asda stores over Christmas period 2014. Collecting customer opinions can be expensive and time consuming task using conventional methods such as surveys. The sentiment analysis of the customer opinions makes it easier for businesses to understand their competitive value in a changing market and to understand their customer views about their products and services, which also provide an insight into future marketing strategies and decision making policies.

General Terms
Natural Language Processing, Opinion Mining.

Keywords
Text Mining, Sentiment Analysis, Open Source, Twitter Data Analysis, Social Data Mining, R Packages.

1. INTRODUCTION
Social media have become an emerging phenomenon due to the huge and rapid advances in information technology. People are using social media on daily basis to communicate their opinions with each other about wide variety of subjects, products and services, which has made it a rich resource for text mining and sentiment analysis. Social media communications include Facebook, twitter, and many others. Twitter is one of the most widely used social media sites. Figure 1, shows mapping the number of twitter messages sent per second worldwide. In the literature, there is no standard method for mining and analyzing social media business data. Here, an open source approach for text mining and sentiment analysis using a set of R packages [2, 6, and 7] for mining twitter data and sentiment analysis is presented, which is applicable for other social media sites. A case study of two UK stores is presented to show the importance of analyzing user generated online opinions from Microblogs. This is helpful for business their performance monitoring from customer perspective instead of making customer surveys, which are expensive and time consuming.

The rest of the paper is organized as follows: Section 2 presents briefly Text mining. Section 3 presents an overview of Sentiment analysis field and related work. Section 4 shows the proposed methodology for mining and sentiment analysis over twitter Microblogs. Section 5 provides the experimental details and shows the results. Section 6 presents the conclusions and implications of the research and also shows future research possibilities.

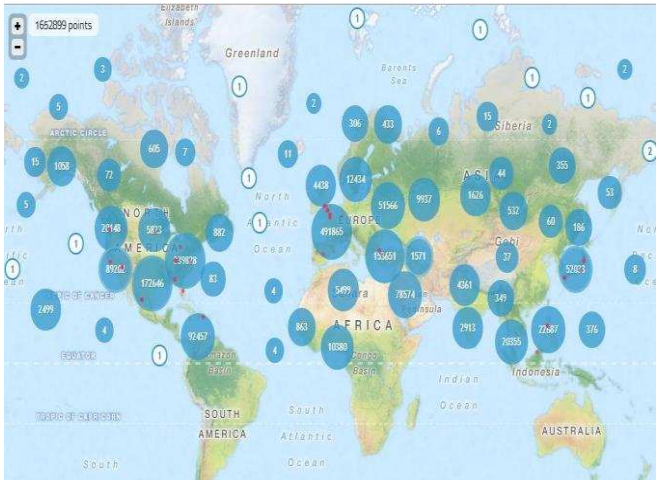

Fig 1: Mapping the number of tweets per second worldwide. As captured from [8].

2. TEXT MINING
Text Mining is the automated process of detecting and revealing new, uncovered knowledge and inter-relationships and patterns in unstructured textual data resources. Text mining targets un-discovered knowledge in huge amounts of text. Whereas, search engines and Information Retrieval (IR) systems have specific search target such as search query or keywords and return related documents [1]. This research field utilizes data mining algorithms, such as classification, clustering, association rules, and many more in exploring and discovering new information and relationships in textual sources. It is an inter-disciplinary research field combining information retrieval, data mining, machine learning, statistics and computational linguistics [1]. Figure 2, summarizes the text mining process. Firstly, a set of un-structured text documents is collected. Then, the pre-processing for the documents is performed to remove noise and commonly used words, stop words, stemming. This process produces a structured representation of the documents known as Term-document matrix, in which, every column represents a document and every row represents a term occurrence throughout the document. The final step is applying data mining techniques such as clustering, classification,

association rules to discover term associations and patterns in the text and then, finally, visualizing these patterns using tools such as word-cloud or tag-cloud.

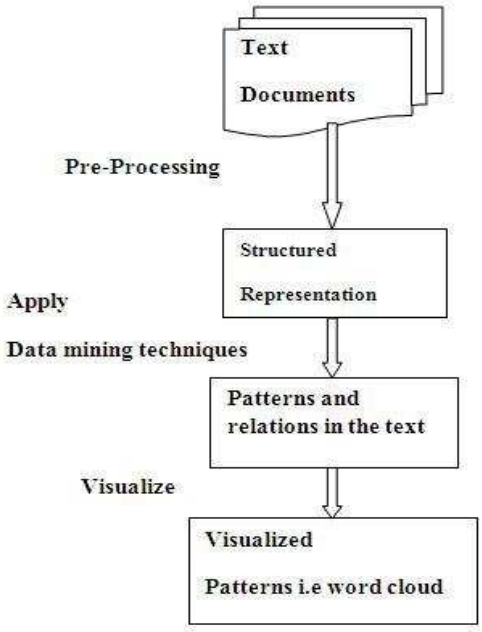

Fig 2: Text Mining Process.

3. SENTIMENT ANALYSIS

Sentiment analysis has been first introduced by Liu, B [4]. It is also known as opinion mining and subjectivity analysis is the process to determine the attitude or polarity of opinions or reviews written by humans to rate products or services. Sentiment analysis can be applied on any textual form of opinions such as blogs, reviews and Microblogs. Microblogs are those small text messages such as tweets, a short message that cannot exceed 149 characters. These microblogs are easier than other forms of opinions for sentiment analysis [11]. Sentiment analysis can be done on a document level or a sentence level. In the first case, the whole document is evaluated to determine the opinion polarity, where, the features describing the product/service should be extracted first. Whereas, the second one, the document is divided into sentences each one is evaluated separately to determine the opinion polarity [4].

3.1 Related Work

Twitter has been used for sentiment analysis in many studies [5, 10, 11, 12, 13, 14 and 21] for different purposes. For example, Zhou et al. and Tumasjan et.al [10, 13], proposed a method for mining opinions from twitter about presidential elections candidates and predicting the election results. Asur et al. [12], provided a model to predict the expected revenues for a movie by analyzing social media data. Das et al. [17], implemented a framework for mining public twitter opinion about Samsung Galaxy phones. Similarly, Mostafa, M. M. [21], used twitter data to evaluate the sentiment of the consumers of big brands such as Nokia, Samsung, IBM and airlines such as Egypt air. Sentiment analysis has been recently applied to many other areas to analyze and predict the public behavior and feelings towards various products, services, social and political events. Sentiment analysis on its own has been applied in [20], to explore the emotions in e-mails and story books.

Sentiment analysis can be performed using two methods. The first is opinion lexicon-based approach [14], in which, the lexicon is composed of a set of positive and negative opinion words, used to score the opinion sentences either, positive, negative or neutral. This approach is very popular and requires a scoring function to score every sentence according to the existence of positive or negative words. Figure 3, shows the lexicon based sentiment analysis method. The lexicon based method uses a lexicon, a set of positive and negative words, combined with a scoring function to determine the sentiment polarity. The second approach is using machine learning techniques for training a classifier using a set of pre-classifies opinions as a training set. Then, use the trained classifier to classify new opinions as positive, negative or neutral [15]. Pak, A et.al [5], used supervised technique to build a classifier using Part of speech tagger and N-gram methods and used the classifier to classify opinions. Research confirmed that lexicon-based methods outperformed machine learning methods [15]. This work utilises lexicon-based method for sentiment analysis.

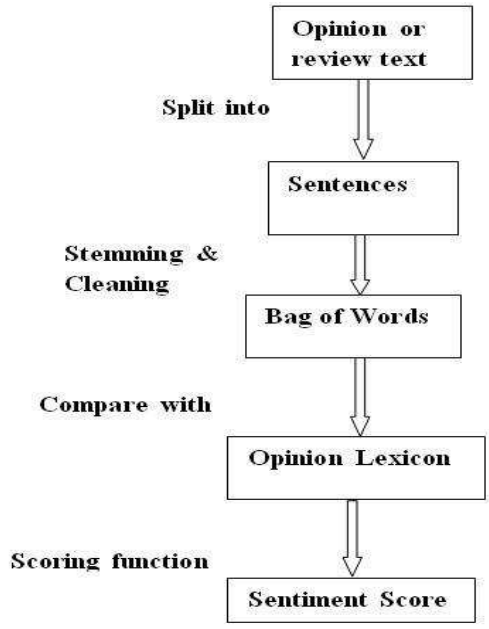

Fig 3: Lexicon-based sentiment analysis approach.

4. THE PROPOSED METHODOLOGY

The methodology used for mining twitter Microblogs is shown in Figure 4. The steps involved in the methodology are described as follows:

1. Data Access: Using TwitteR package to make a keyword search to access twitter messages.

2. Data Cleaning: Using some additional tm package to get the tweet text, then, clean the data from stop words (non-functional), removing spaces, punctuation, URLs and performing stemming (get the root of the words). This step produces a structured representation of tweets called Term-Document Matrix.

3. Data Analysis: The structured representation produced in the previous step enables performing Mining tasks such as finding association rules, finding more frequent terms and performing sentiment analysis using the lexicon-based approach, which uses a set of positive and negative words. A scoring function is used to assign a score for each tweet.

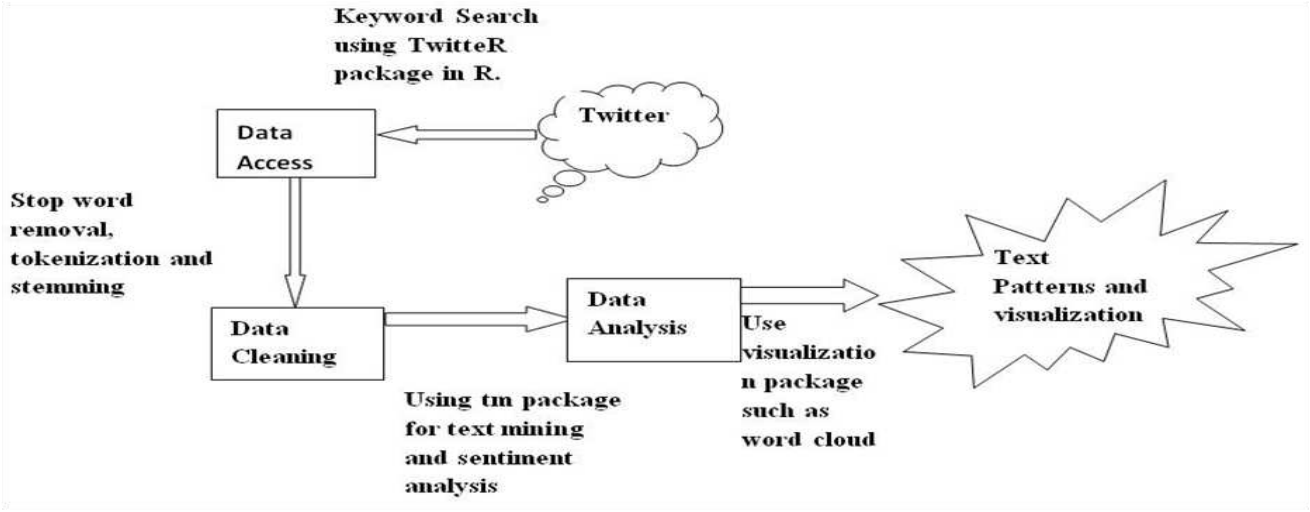

Fig 4: The proposed Methodology for text mining and sentiment analysis for Twitter Microblogs.

4. Visualization: The wordcloud package [18] and bar plots has been used to show the frequency of words in the customer tweets and the sentiment scores.

5. EXPERIMENTS AND RESULTS

The experiment involved collecting 2000 tweets for Tesco and 803 tweets for Asda over Christmas period 2014. The twitteR [7], R package has been used to access twitter data. The package enables authentication and access to twitter messages by using keyword search queries [9]. After getting the data, the extraction of tweets text and the cleaning has been done using tm package [19]. The output obtained from the previous step is a structured representation of tweet text, tweet-term-matrix. This structured representation cab be used to perform text mining using tm package. One of the outputs of this is the word cloud representation of tweets. This cloud showed in figures 5 and 6 for Tesco and Asda respectively. The size of each term in the cloud indicates the number of mentions of that term in the tweets, reflecting its importance. It is also possible to obtain the most frequent terms with a specific occurrence threshold.

In addition, it is also possible to compute the term associations. For example, we can find the associations between the term "offers" and the rest of the matrix. This shows the most and least interesting offers for customers.

Related to sentiment analysis, it has been done utilising the lexicon-based approach shown in Figure 3. We now have the bag of words representation of tweets after applying text mining. For us to perform the sentiment analysis, we need the opinion lexicon and a scoring function, to assign scores to tweets. We used the lexicon available in [3] and the scoring function is shown in equation 1.

Sentiment Score= ∑positive words – ∑ Negative words

→… (Equation 1)

The score can be 0, if the positive and negative words are equal or there is no existence of any opinion words in the text. (Neutral).

The score can be positive, if the number of positive words > the number of negative words. (Positive polarity)

The score can be negative, if the number of negative words > the number of positive words. (Negative polarity)

The results of the sentiment analysis are shown if Table 1, showing the sentiment scores and frequency of tweets having that score for both retailers Tesco and Asda. Moreover the mean sentiment score for Tesco is 0.1595 and for Asda is -0.00373599 and the Media for both is 0.0. Figures 7 and 8 show the plots of sentiment scores obtained.

Table 2, shows some examples of tweets. It is clear that the first 4 tweets have positive polarity. Whereas, the last tweet has a negative polarity.

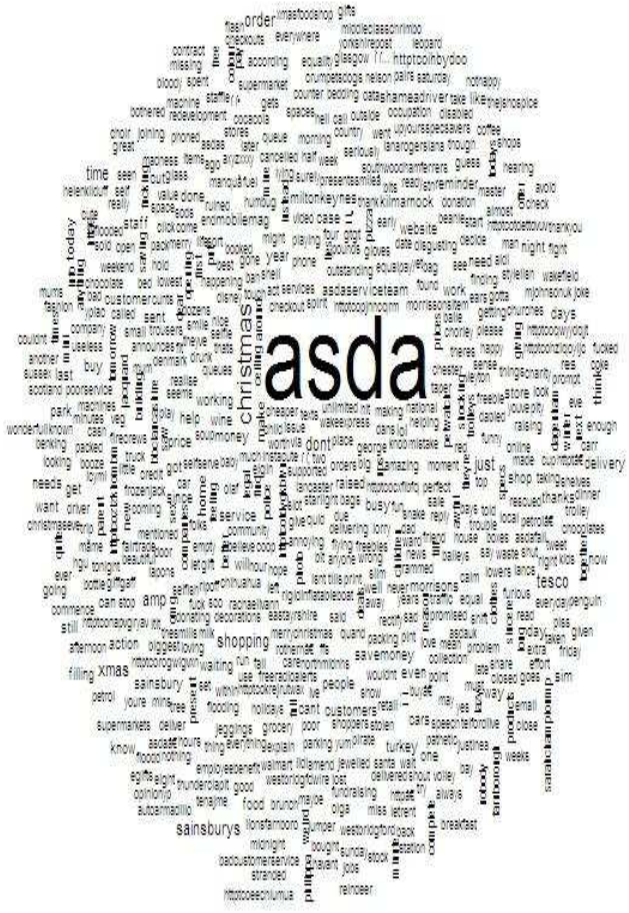

Fig 5: Asda tweets word-cloud.

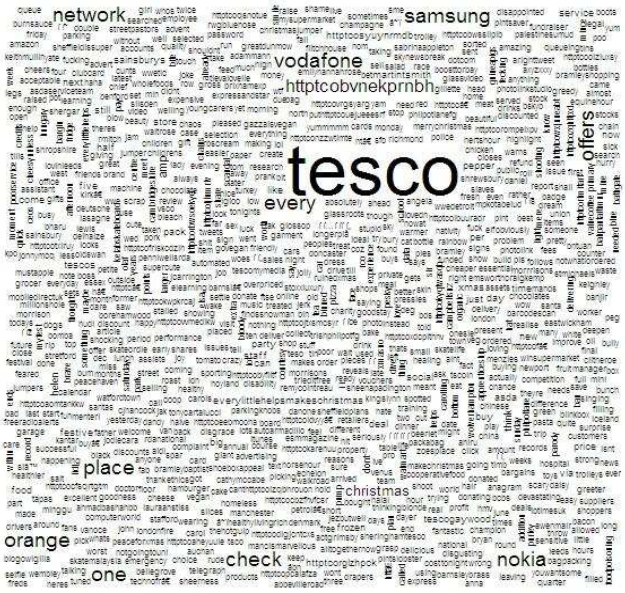

Figure 6: Tesco tweets word-cloud.

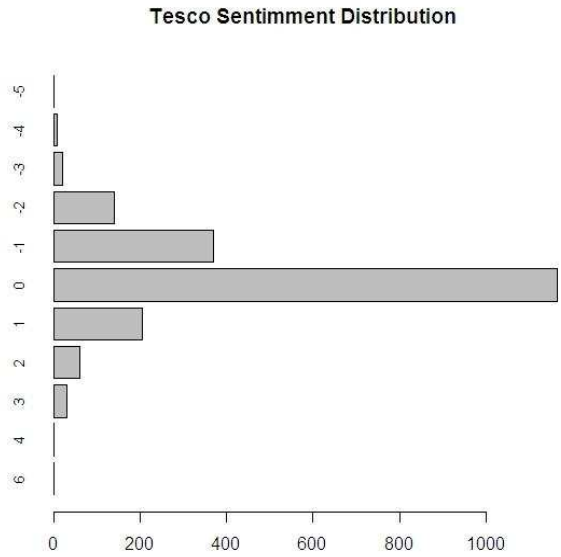

Fig 7: Sentiment Score distribution for Tesco.

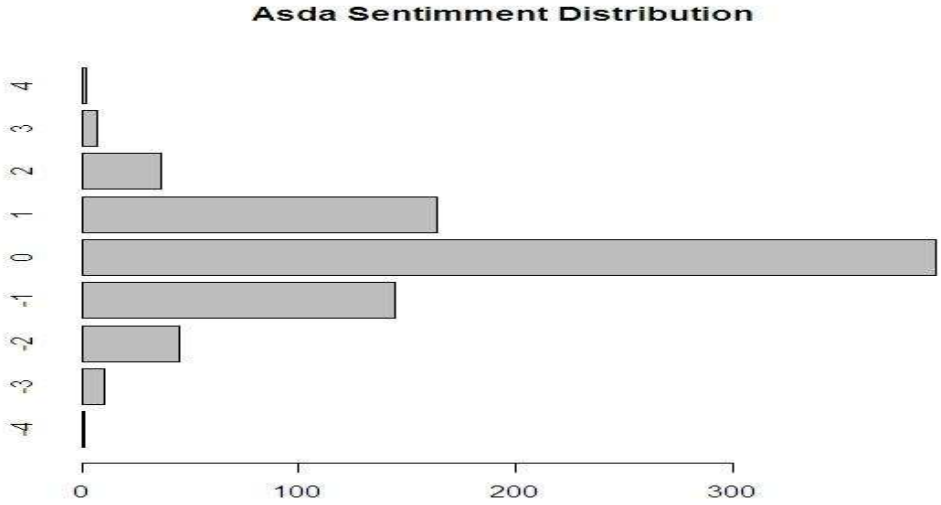

Fig 8: Asda tweets Sentiment distribution.

Table 1: Showing Sentiment Polarity, Scores and their Frequency for Tesco and Asda data.

		Negative					Neutral	Positive				
	Sentiment Score	-5	-4	-3	-2	-1	0	1	2	3	4	6
TESCO	Frequency	1	1	31	59	205	1166	369	139	21	7	1
ASDA	Frequency	0	1	10	45	144	394	164	36	7	2	0

Table 2: Showing examples of positive and negative tweets.

"Had the worst ever customer service experience #tesco Stretford
Tesco Very good of you to donate to the school"
"Happy to see #Tesco fight back at ground level. Keep at it!"
"Asda in Grangemouth, have a fantastic display #christmas #asda"
"That annoying moment when @asda overcharge you #asda"

6. DISCUSSIONAND CONCLUSIONS

1. It has been confirmed by this investigation that, it is possible to collect, pre-process, analyse and visualize twitter social data using R statistical software open source tool packages.

2. It is viable to apply text mining tasks and sentiment analysis for twitter data to analyse user contributed reviews for products or services.

3. It will provide a competitive advantage for business retailers and service providers to analyse their customer views regarding their product or service using social media data. This will help them improve their business value and better manage their customer relationship.

4. The described approach is applicable on other social media data sources such as Facebook.

It can be generalized that, Businesses can utilize their consumer opinions generated from social media tracking and analysis by adapting their marketing plans, products and business intelligence respectively. An important perspective for future work could be building social media tracking and monitoring system as opinions are changing over time. Moreover, it is also valuable to use un-supervised techniques in sentiment analysis and opinion mining for improving the business competitive value and the customer relationship management. In addition to comparing various sentiment classification techniques utilised for opinion mining. What is more social media can be used as a tool for sales prediction using opinion mining.

7. REFERENCES

[1] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

[2] Feinerer, I. (2014). Introduction to the tm Package Text Mining in R. nd): n. pag. Web.

[3] Jeffrey Breen, https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English. Accessed: 26th of December 2014.

[4] Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2, 627-666.

[5] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

[6] http://www.r-project.org/. Accessed: 26th of December 2014.

[7] http://cran.r-project.org/web/packages/twitteR/index.html. Accessed : 26th of December 2014.

[8] http://onemilliontweetmap.com/. Accessed : 27th of December 2014.

[9] Danneman, N., & Heimann, R. (2014). Social Media Mining with R. Packt Publishing Ltd.

[10] Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013, June). Sentiment analysis on tweets for social events. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE.

[11] Milstein, S., Lorica, B., Magoulas, R., Hochmuth, G., Chowdhury, A., & O'Reilly, T. (2008). Twitter and the micro-messaging revolution: Communication, connections, and immediacy--140 characters at a time. O'Reilly Media, Incorporated.

[12] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

[13] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

[14] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[15] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011, 89.

[16] Bhuta, S., & Doshi, U. (2014, February). A review of techniques for sentiment analysis Of Twitter data. In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on (pp. 583-591). IEEE.

[17] Das, T. K., Acharjya, D. P., & Patra, M. R. (2014, January). Opinion mining about a product by analyzing public tweets in Twitter. In Computer Communication and Informatics (ICCCI), 2014 International Conference on (pp. 1-4). IEEE.

[18] Ian Fellows, http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf.

[19] Feinerer, I., & Hornik, K. (2012). tm: Text Mining Package. R package version 0.5-7.1.

[20] S. Mohammad (2012), "From once upon a time to happily ever after: Tracking emotions in mail and books", Decision Support Systems, 53 (2012), pp. 730–741.

[21] Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications, 40(10), 4241-4251.