ETL Transformation Algorithm for Facebook Opinion Data

Afef Walha[1], Faiza Ghozzi[1,2] and Faïez Gargouri[1,2]
[1]*Multimedia, InfoRmation Systems and Advanced Computing Laboratory, Sfax, Tunisia*
[2]*Institute of Computer Science and Multimedia, Sfax University, Sfax, Tunisia*

Keywords: ETL, Sentiment Analysis, Social Network.

Abstract: Considered as a rich source of information, social networking sites have been created lot of buzz because people share and discuss their opinions freely. Sentiment analysis is used for knowing voice or response of crowd for products, services, organizations, individuals, events, etc. Due to their importance, people opinions are analyzed in several domains including information retrieval, semantic web, text mining. These researches define new classification techniques to assign positive or negative opinion. Decisional systems like WeBhouse, known by their data-consuming must be enriched by this kind of pertinent opinions to give better help to decision makers. Nevertheless, cleaning and transformation processes recognized by several approaches as a key of WeBhouse development, don't deal with sentiment analysis. To fulfill this gap, we propose a new analysis algorithm which determines user's sentiment score of a post shared on the social network Facebook. This algorithm analyzes user's opinion depending on opinion terms and emoticons included in his comments. This algorithm is integrated in transformation process of ETL approach.

1 INTRODUCTION

Sentiment analysis is concerned with the automatic extraction of sentiment-related information from text. Most sentiment analysis addresses commercial tasks, such as extracting opinions from product reviews. People can now post reviews of products at merchant sites and express their views on almost anything in social Websites.

With the growing availability and popularity of opinion-rich resources such as social networking sites (e.g. Facebook, Twitter), new opportunities and challenges arise. In these sites, millions of users interact frequently and share variety of digital content with each other. They express their feelings and opinions on every topic of interest. These opinions carry import value for personal, academic and commercial applications. Social networking sites represent new and measurable sources of information to an organization, such as customer's opinions on some products. These opinions may be helpful for decision making.

Existing opinion analysis approaches propose classification techniques and methods in order to detect sentiment polarity. These approaches cover many research domains including information retrieval, text mining and semantic web. Decisional systems can adopt proposed classification techniques in order to integrate opinion analysis in Data WeBhouse (DWB). Nevertheless, ETL design is recognized as complex task. It is more and more difficult including these techniques to analyze pertinent opinion data.

In our previous works, we proposed an ETL processes design approach integrating user's opinion available on Facebook social network. This approach offers generic ETL operators to Webhouse designer reducing the complexity of tackling opinion extraction and transformation from Facebook source. In this paper, we focus on opinion analysis step of ETL transformation process which adapts a lexicon sentiment analysis method. We propose an algorithm that determines the user's sentiment score reflecting his opinion about a product or service shared on Facebook pages. This score is resulted by analyzing user's comments based on lexical DB composed of emoticons and opinion words dictionaries.

This paper is organized as follow: section 2 presents a brief review on ETL design and opinion analysis approaches. Then, we present an overview of our ETL design approach integrating sentiment analysis. In section 4, transformation process is enriched by a new algorithm that combines visual

Walha A., Ghozzi F. and Gargouri F..
ETL Transformation Algorithm for Facebook Opinion Data.
DOI: 10.5220/0005494101500155
In *Proceedings of the 11th International Conference on Web Information Systems and Technologies* (WEBIST-2015), pages 150-155
ISBN: 978-989-758-106-9

cues (emoticons) and opinion words collected from user's comments to determine his opinion polarity. Finally, we conclude and present some perspectives in section 5.

2 RELATED WORK

2.1 ETL Modelling Approaches

ETL processes design is a crucial task in DW development due to its complexity and its time consuming. Works dealing with this task can be classified into two main groups: *Specific ETL modelling* and *Standard ETL modelling*. The first group offers specific notations and concepts to give rise for new specialized modelling languages. ETL processes proposed in (Vassiliadis, 2009) are limited to typical activities (e.g. join, filter). (El-Sappagh et al, 2011) extend these proposals by modelling advanced operations, like user defined functions and conversion into structure, etc. In order to design complex ETL scenario, specific modelling approaches propose conceptual or formal models. However, the standardization is an essential asset in modelling. The goal of the second group is to overcome this problem by using standard languages like UML, BPMN, etc. (Trujillo and Luján-Mora, 2003) and (Muñoz et al, 2010) use UML class diagram to represent ETL processes statically or dynamically by using UML activity diagram. (Wilkinson et al, 2010) and (Akkaoui et al, 2012) use BPMN standard where ETL processes can be a particular type of business process.

Even though ETL modelling approaches succeeded in providing interesting several modelling methods and techniques, they don't cover pertinent opinion data sources available on web sources like social networks, blogs, reviews, etc.

2.2 Opinion Analysis Approaches

Opinions are usually subjective expressions that describe people sentiments, appraisals or feelings toward entities, events and their properties.

Integrating opinion data is nowadays a hot topic for many researchers. The common goal of sentiment analysis approaches is to detect text polarity: positive, negative or neutral. (Medhat et al, 2014) categorize opinion analysis approaches into **machine learning** and **lexicon approaches**. *Machine learning* approaches ((Wilson et al, 2005), (Abbasi et al, 2008)) use classification techniques (e.g. Naive Bayes, maximum entropy, and support vector machines). *Lexicon approaches* rely on a sentiment lexicon, a collection of known and precompiled opinion terms. They use sentiment dictionaries with opinion words and match them with data to determine text polarity. They assign sentiment scores to opinion words according to positive or negative terms contained in the dictionary. Lexicon approaches are divided into *dictionary-based approaches* and *corpus-based approaches*.

Dictionary-based approach ((Kim and Hovy, 2004), (Hu and Liu, 2004)) begins with a predefined dictionary of positive and negative words, and then uses word counts or other measures of word incidence and frequency to score all opinions in the data. The idea of these approaches is to first manually collect a small set of opinion words with known orientations (seed list), and then to grow this set by searching in a known lexical DBs (e.g. WordNet dictionary) for their synonyms and antonyms. The newly found words are added to the seed list (Liu, 2011). Opinion words share the same orientation as their synonyms and opposite orientations as their antonyms. (Qiu et al, 2010) and (Hu and Li, 2011) use this technique to find semantic orientation for adjectives. (Qiu et al, 2010) worked on web forums to identify sentiment sentences in contextual advertising.

Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based method can produce opinion words with relatively high accuracy. This method needs very large labeled training data. (Jiao and Zhou, 2011) use Conditional Random Fields methods in order to discriminate sentiment polarity by multi-string pattern matching algorithm applied on Chinese online reviews in order to identify sentiment polarity. They established emotional and opinion words dictionaries.

Machine learning and lexicon approaches use opinion words and classification techniques to determine text polarity. In addition to the use of opinion words to analyze sentiment, emoticons decorating a text can give a correct insight of the sentence or text. For example, the emoticon "☺" expressing "*happiness*" means positive opinion. Further researchers take care of the increasing using of these typographical symbols for sentiment classification. In (Vashisht and Thakur, 2014), authors identify the possible set of emoticons majorly used by people on Facebook and use them to classify the sentiment. Then, they use a finite state machine to find out the polarity of the sentence or paragraph. The problem with this approach is performing sentiment analysis on text-based status

updates and comments, disregarding all verbal information and using only emoticons to detect both positive and negative opinions. (Hogenboom et al, 2013) propose a framework for automated sentiment analysis, which takes into account information conveyed by emoticons. The goal of this framework is to detect emoticons, determine their sentiment, and assign the associated sentiment to the affected text in order to correctly classify the polarity of natural language text as either positive or negative.

Existing ETL design approaches model various web sources without considering user opinions available on these sources including social networks, reviews, blogs, forums or emails, etc. In the past few years, many researchers have shown interest to opinions expressed by people on any topic. They proposed sentiment analysis methods and techniques to determine text polarity. Some approaches apply classification algorithms and use linguistic features (machine learning approaches). Others use sentiment dictionaries with opinion words and match them with data sources to determine text polarity (*lexicon approaches*). These approaches assign sentiment scores to text according to positive or negative words contained in the dictionary. Others researchers use emoticons to disambiguate sentiment when it is not conveyed by any clearly positive or negative words in a text segment.

Sentiment analysis approaches presented in the literature are very helpful and interesting to classify text polarity. In spite of the importance of sentiment classification approaches, we note that few of them employ the coupling between sentiment analysis and ETL processes in order to enhance semantic orientation to multidimensional design. We propose an ETL design approach adopting lexicon sentiment analysis method. We consider Facebook opinion data as a source to ETL processes. In the current work, we define a new algorithm that analyzes user's comments about a product described on a Facebook post and assign a sentiment score to him. This score reflects user's opinion. It is determined based on emoticons and opinion words polarities defined on lexical DB dictionaries.

3 ETL DESIGN APPROACH OVERVIEW

In (Walha et al, 2015), we define a new ETL design approach that integrates people's opinions to model **Extraction**, **Transformation** and **Loading** processes. Figure 1 shows an overview of this approach.

Extraction process starts by collecting general

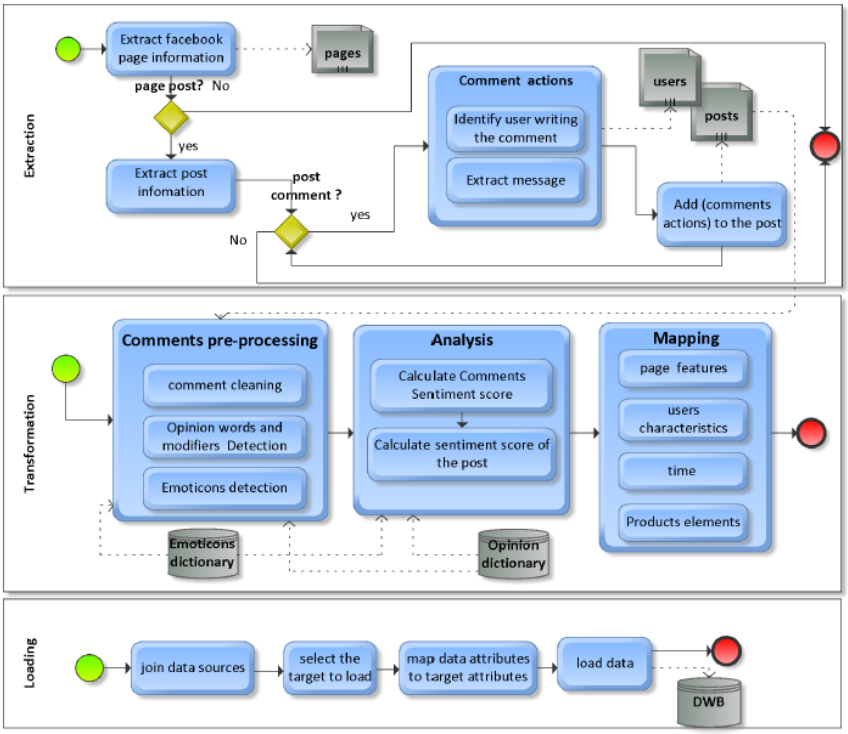

Figure 1: Overview of ETL design approach.

information about a post shared on a Facebook page. A post is an individual entry of a user, page, or group about a product or service. A list of available actions (likes and comments) can be associated to each post. These actions can help to gather people opinions a post. The next step consists in identifying users' comments associated with the post.

Transformation process is organized in three main steps: ***pre-processing***, ***analysis*** and ***mapping***. ***Pre-processing*** involves comments cleaning and comments' opinion words and emoticons detection. ***Analysis*** is the main step of transformation process. It consists on analyzing user's comments realized on the post in order to classify his opinion (positive or negative) about the product described in this post. We propose for that an algorithm (section 4.1) which assigns to each post a user's sentiment score. This proposal analyzes verbal cues (opinion words) and visual cues (emoticons) contained in the post's comments. This algorithm adopts a lexicon sentiment analysis method. It is based on emoticons and opinion dictionaries defined in the lexical DB (Walha et al, 2015). In these dictionaries, we associate for each opinion term and emoticon a sentiment polarity score which can be positive (between (0) and (1)) or negative (between (-1) and (0)). For examples, the polarity (0.9) is associated to the opinion word "excellent" expressing "Happiness" sentiment and the polarity (-0.7) corresponds to the emoticon "☹" that expresses the "sadness". The final step in transformation process is the ***mapping***. It aims to match between ETL source (concepts of "Facebook" model) and the target (DWB model).

Loading process feeds the DWB with data resulted from transformation step. It consists in loading data into DWB multidimensional elements including dimensions, measures, facts, attributes and parameters.

4 TRANSFORMATION PROCESS

4.1 Opinion Analysis Algorithm

Transformation process analysis step (figure 1) aims to detect users' opinion according to their comments. The goal of ***PostSentimentScore*** (algorithm 1) is to determine a sentiment score *(SentP)* to a post *(P)*. *SentP* reflects opinion of the user *(U)* about the product described in the post (P). The user *(U)* can have a positive opinion if *SentP* is comprised between (0) and (1), or negative opinion with a value comprised between (-1) and (0). The principle of algorithm 1 is to detect comments realized by the user *(U)* on the post *(P)* and then calculate their average, which corresponds to the post sentiment score *(SentP)*.

Algorithm 1: PostSentimentScore.

Input : P *// A post shared on a Facebook page.*
U *// User who comments P.*
Output : SentP *// sentiment score assigned to the post P*
1: SentP ← 0
2: N ← **count** (comments) // numbers of comments published by the user (U) on the post (P)
3: **For each** Ci associated to the post P shared by U **do**
SentP ←SentP+**CommentSentimentScore (Ci)**
EndFor
3: **return** SentP / N

Algorithm 2: CommentSentimentScore.

Input : C // A comment shared by the user (U) on (P)
Output : SentC // sentiment score of the comment (C)
1: w ← **countOpinionWords** (C) // number of opinion words in C
2: e ← **countEmoticons**(C))// number of emoticons in C
3: SentC ← 0
4: **For each** wj of the comment C **do**
SentWj ← **getOpinionWordPolarity** (wj) *// get the polarity of opinion word wj defined in opinion dictionary*
If modifier (wj) = true **then**
SentMj ← **getModifierPolarity** (mj) *// get the polarity of the modifier mj defined in opinion dictionary*
If SentMj > 0 **then**
S ← 1
Else
S ← -1
endIf
Else
SentMj ← 0
EndIf
SentC ← SentC + S * (|SentMj| + SentWj) / 2
EndFor
5: **For each** ej of the comment C **do**
SentEj ← **getEmoticonPolarity** (ej) *// get the polarity of emoticon ej defined in emoticon dictionary*
SentC ← SentC + SentEj
EndFor
6: **If** w + e > 0 **then**
SentCi ← SentCi / (w + e)
endIf
7: **return** SentC

The score of the comment *(C)* is determined by algorithm 2, untitled ***CommentSentimentScore***. Its principle is the following. First, it computes the numbers of emoticons *(e)* and opinion words *(w)* contained in *(C)*. Then, it initializes *SentC*, i.e. sentiment score of the comment *(C)*, to the value (0). This score is increased by polarity scores of all emoticons and opinion words used in *(C)*. These scores are defined in the lexical DB (emoticon and

opinion dictionaries). Comment opinion words can be related to a modifier, which can change its sentiment polarity (e.g. the modifier "not" in the comment "not good" change the user's opinion). For that, we verify the existence of modifier *(mj)* related to each opinion word *(wi)* used in *(C)*. A modifier *(mj)* may change the polarity of *(wj)*. For that, we define the variable *(S)*. Its value depends on modifier polarity score *(SentMj)*. It is equal to (1) in case of positive value of *(SentMj)*. Otherwise, the value (-1) is associated to *(S)*. Comment's sentiment score *(SentC)* is added to the average of modifier's polarity absolute value *(SentMj)* and opinion word's polarity *(SentWj)* multiplied by *(S)*.

In our approach, we combine the use of opinion terms and emoticons to detect user's opinion expressed on a comment. *(SentC)* is then increased by the sum of emoticons polarities and finally divided by the sum of emoticons *(e)* and opinion words *(w)* used in *(C)*.

4.2 Transformation Prototype

Facebook data are collected through Facebook Graph API Explorer tool (API, 2015). To integrate user's opinion in the ETL prototype, we use this tool to extract information about a Facebook post, including post name, message, created_time, pageName, link, type, etc. Also, we obtain user's comments realized on a post. Data collected from Facebook about posts and their users' comments are converted into XML files including USERS, PAGES, PRODUCTS, and POSTS. These files collection composes our XML source DB.

The main goal of our ETL transformation process is to analyze user's opinion through (**transformation sentiment analysis** step). We adopt a lexicon based opinion analysis method. We propose for that a lexical DB composed of emoticons and opinion dictionaries. These latter are transformed into XML files containing opinion word, modifier and emoticon, their associated sentiment classes (e.g. happiness) and polarity scores. Figure 2 (XML lexical DB) depicts three lexical DB files "*emoticonsSample.XML*", "*opinionWordsSample.XML*" and "*modifiersSample.XML*".

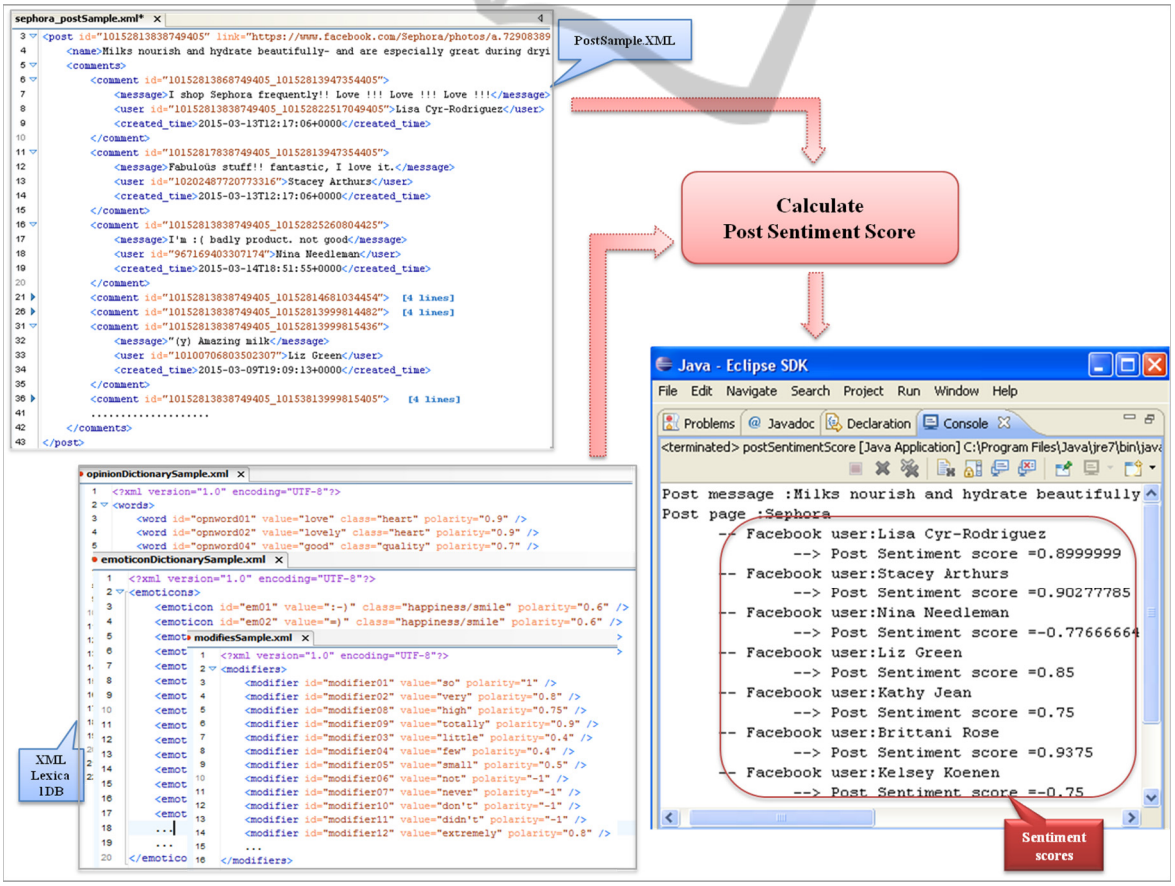

Figure 2: Transformation prototype.

In figure 2, a sample of post (P) is also presented in "*PostSample.XML*" file. To detect users' opinions based on their comments on the post (P), we apply **PostSentimentScore** (algorithm 1) which returns a sentiment score for each user that comments (P). This score depends on opinion words and emoticons exploited in comments. For that, we apply **CommentSentimentScore** (algorithm 2). Results are depicted in figure 2 (*Post Sentiment Scores*).

5 CONCLUSION AND FUTURE WORKS

Due to the importance of people's opinions expressed on social networks for decisional systems, we worked on integrating them in ETL processes design. In this paper, we focus on ETL **transformation** process. We propose a new algorithm which analyzes user's opinions expressed through comments about a post shared on the social network Facebook. Its goal is to detect both positive and negative polarity. We associate for that a sentiment score depending on comment's opinion terms and emoticons. In the proposed algorithm, sentiment analysis adopts a lexicon method based on opinion and emoticons dictionaries.

As future works, we intend to enrich our lexical DB in order to adapt context-specific opinion analysis. Also, we will extend our ETL processes design approach by integrating more opinion web sources including clickstreams, web sites, and others social networks.

REFERENCES

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. In *ACM Transactions on Information Systems Journal*.

Akkaoui, Z., E., Mazón, J., Vaisman, A. A., Zimányi, E., 2012. BPMN-Based Conceptual Modeling of ETL Processes. In *DAWAK'12, 14th International Conference on Data Warehousing and Knowledge Discovery*, pages 1-14, Springer.

API, 2015. API Graph Explorer Tool, "https:// developers.facebook.com/tools/explorer".

El-Sappagh, S., H., Hendawi, A., M., Bastawissy, A., H., 2011. A proposed model for data warehouse ETL processes. *In Journal of King Saud University - Computer and Information Sciences,* pages 91-104, Elsevier.

Hogenboom, A., Bal, D., Frasincar, F., 2013. Exploiting Emoticons in Sentiment Analysis. In *SAC'13, 28th Annual ACM Symposium on Applied Computing*, pages 703-710.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In *KDD'04, international conference on Knowledge Discovery and Data Mining*, pages 168-177, ACM.

Hu, Y., Li, W., 2011. Document Sentiment Classification by Exploring Description Model of Topical Terms. In *Computer Speech Language Journal*, pages 386-403, Elsevier.

Jiao, J., Zhou, Y., 2011. Sentiment Polarity Analysis based Multi Dictionary. In *ICPST'11, International Conference on Physics Science and Technology,* Elsevier.

Kim, S., Hovy, E., 2004. Determining the Sentiment of Opinions. In *COLING'04, 20th International conference on Computational Linguistics*.

Liu, B., 2011. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag Berlin Heidelberg, 2nd Edition.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment Analysis Algorithms and Applications: A Survey. In *Ain Shams Engineering Journal,* pages 1093-1113.

Muñoz, L., Mazón, J.N., Trujillo, J., 2010. A Family of Experiments to Validate Measures for UML Activity Diagrams of ETL Processes in Data Warehouse. In *Information & Software Technology*, pages 1188-1203, Elsevier.

Qiu, G., He, X., Zhang. F., Shi, Y., Bu, J., Chen, C., 2010. DASA: Dissatisfaction-Oriented Advertising Based on Sentiment Analysis. In *Expert Systems with Application Journal*, pages 6182–6191, Elsevier.

Trujillo, J., Luján-Mora, S., 2003. A UML Based Approach For Modeling ETL Processes in Data Warehouses. In *ER'03, 22nd International Conference on Conceptual Modeling,* pages 307-320, Springer.

Vashisht, S., Thakur, S., 2014. Facebook as a Corpus for Emoticons-Based Sentiment Analysis. In *IJETAE'14, International Journal of Emerging Technology and Advanced Engineering*, pages 904-908.

Vassiliadis, P., 2009. A Survey of Extract–Transform–Load Technology. In *IJDWM'09, International Journal of Data Warehousing & Mining,* pages 1-27.

Walha, A., Ghozzi, F., Gargouri, F., 2015. ETL design toward social network opinion analysis. In SERA'15, *13th IEEE/ACIS on Software Engineering, Reasearch, Management and applications,* Springer (to appear).

Wilkinson, K., Simitsis, A., Dayal, U., Castellanos, M., 2010. Leveraging Business Process Models for ETL Design. In *ER'10, 29th International Conference on Conceptual Modeling,* Springer.

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *HLT'05, 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pages 347-354.