Data Mining through Sentiment Analysis: Lexicon based Sentiment Analysis Model using Aspect Catalogue

Aman Mehto
Computer Science Department, Medi-caps Institute of Technology and Management
Indore Madhya Pradesh, India - 453331
aman.gijoe@gmail.com

Karnika Indras
Computer Science Department, Medi-caps Institute of Technology and Management
Indore Madhya Pradesh, India - 453331
karnika.indras6@gmail.com

Abstract— **Keeping in mind the peculiarity about product reviews, we have proposed a method that can help in determining the sentiment associated with the product in any review. Our method is based on 'Lexicon based approach for Sentiment Analysis' with an additional feature of Aspect Catalogue. In our approach we first look for the keywords present in aspect catalogue to identify those sentences in which features of any product are mentioned. Taking into account only these sentences we further work on determining what opinion the speaker holds about features or aspects of the product. Aspect catalogue is referred again to find degree of importance corresponding to the feature with respect to the product. This is followed by calculation of numerical value of sentiment involved by taking the average of all the weightage of the features present in the text.**

Keywords— Aspect Catalogue; Aspect Weightage; Aspect Polarity; Lexicon; Product Sentiment; Product Event; Integration Algorithm;

I. INTRODUCTION

With the surge in the opinion rich resources on internet such as discussion forums, peer to peer networks, blogs, reviewing websites etc. has given us opportunity to get insights into the perspective of the people or communities towards any service or any product. These reviews are like gold mines to the companies. As these data are active feedback which can help companies developing the product, identify what is being liked or disliked about the product. These can further help implementation of recommendation system by identifying the choices and preferences of users.

In the specific case of product reviews, companies are mainly interested in identifying and gathering opinions related to the features of the product. This means that they generally have a focused demand of extracting the perception of user about its features. For example consider a mobile phone review "the display quality is great, the icons pop out on the screen and watching pictures and movies on it is a treat but the only thing I am not happy is the flash, it is not as powerful as it should be". The whole text of the reviewer is centered on two features of the phone i.e. camera and the flash. Therefore it is wise to give more importance to features associated with any product rather than paying attention at the whole sentence.

In today's revolutionary age of 'Big Data' where the most ordinary form of data can be manipulated to turn into better user experience and hence a commercial profit by the service provider. In this academic paper we have explored different existing methods for sentiment analysis but our main discussion is concentrated to our new model which is based on Lexicon based approach by taking in account the aspects or feature of the subject. We will explore three main implementations in sentiment analysis.

A. Lexicon based Sentiment Analysis Model using Aspect Catalogue

This model will detect and rate a subjective article on the basis of its intended meaning and tell the data model whether the article is intended to positive negative or neutral. In this we will use a two dimensional grid which will analyze a word on its intended meaning and context in which it is said to detect underlying sentiments such as sarcasm satire praise or slander.

B. Basic guidelines for Sentiment Analysis Model (SAM)

During integration of the polarity based sentiment analysis model into the systems(like social media, data mining etc.) this set of guidelines will ensure smooth functioning of the core model while extracting and using all the products obtained from the sentiment analysis model.

C. Applications of SAM (Post Integration)

After integration of the SAM into the core model we will see how it affects the existing model and what the extra features that can be achieved are. We will also see some examples of these core models and see what the differences in the core model pre are and post integration.

II. SENTIMENT ANALYSIS MODEL

A. Lexicon based Sentiment Analysis Model using Aspect Catalogue

- Review of consumer product or services / Review related websites: - this is one of the most common area of

978-1-5090-0669-4/16/$31.00 ©2016 IEEE

application of sentiment analysis. Customer's review about product and services not only aid people to decide which product they should buy but also aids the company to monitor their reputation through these reviews.

- Recommendation Systems: - Recommendation systems may help in making suggestions for products to the users based on their interests. Another application is "flame" detection where highly heated or antagonist language can be identified. These system can also help in providing user with product ads relevant to their needs and interest.
- Business and Government intelligence: - The opinions of the customers can help in business related decisions as well. Tracking of public viewpoint can be helpful in identifying potential customers, there preferences this can further help in maintaining public relations and even in predicting the future trends in business. Similarly, Government intelligence is another area of application where through sentiment analysis the opinion tracking of the people regarding any government service to identify its strength or weakness and in identifying success related to any campaign and many more.

B. Basic Terminologies

In this section we will discuss some basic terminologies which the reader may encounter in the paper.

- Sentiment analysis (bing liu book) :- sentiment analysis is the field of study that analyses people's opinions, sentiments, evaluation, appraisals, emotions and attitudes towards entities such as products , services, individual , organizations , issues , events and their attributes. The research on sentiments and opinions appeared earlier (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000). But the term sentiment analysis particulary appeared first time in (Nasukawa and Yi, 2003), and the term opinion mining first appeared in (Dave, Lawrence and Pennock, 2003). In this paper we have used the term sentiment analysis and opinion mining interchangeable. There have been involvement of different terms in this field with slightly different task such as opinion extraction , sentiment mining , subjective analysis, review mining , emotion analysis as well. But in particular sentiment analysis and opinion mining focus on finding opinion which express positive or negative sentiment.
- Subjectivity and objectivity – An objective sentence is the one which has some factual information. On the other hand subjective sentence is the one in which speaker's emotions, beliefs and views are present. For example -"Delhi is capital of india." Is an objective sentence whereas "I like the city Delhi." is subjective sentence.
- Sentiment Orientation – the evaluative factor of a word is known as its semantic orientation. (Hatzivassiloglou and McKeown [1997]. A positive semantic orientation denotes positive evaluation (i.e. praise) and negative semantic orientation denotes negative evaluation (i.e. criticism). Semantic Orientation has both direction (positive and negative) and intensity (mild or strong). In other words semantic orientation is the numerical treatment of sentiments and opinions. The average semantic orientation of the words present in the sentences of any review can help in classifying the review as positive or negative.

C. Challenges with Sentiment Analysis

- Recognizing fake and spam review – not all the content present on web is genuine and hence this can deviate the results away from the true facts. Therefore preprocessing of web content must include identification of duplicate content, reputation of reviewer (liu 2008), the outliners etc.
- Incorporation of implicit behavior of data – the implicit behavior of data conveys the actual meaning of data. In natural language it's not merely the semantics of the word which are capable of describing the sentiment. for example the word tough For example, in political speech the word "tough" is negative when referring to an economic situation, but becomes positive if the candidate refers to his credentials as a crime-fighter ("tough on crime").
- Domain independence – the sentiment word have different meaning in different domain and hence they pose a problem as one feature set may give good performance in one domain but may have poor performance in some other domain.
- Natural language processing overheads – natural language processing can be daunting at times because natural language contains slang words, ambiguous words, co-reference, implicitness, hindrances etc. These require additional workout to derive the actual meaning of the content.
- Asymmetry in availability of opinion mining software- the software for opinion mining is mostly used by big organizations and government as it is expensive. It is not available to average citizens to take advantage.

D. Existing Models

Existing methods of sentiment analysis can be classified on the basis of the approaches that follow. This section includes the methods which are generally used for extracting polarity of the text.

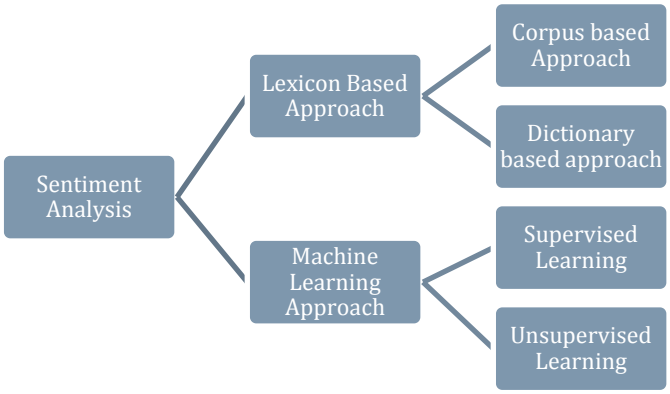

Figure 1.A

- LEXICON BASED METHOD: - This method is based on finding semantic orientation of text i.e. finding positivity, negativity or the neutrality of the text by referring to the lexicon or say dictionaries which identify the positive or negative label in terms of polarity and strength related to their meaning. These dictionaries can be manually created or can be automatically created with the help of seed word. Commonly used lexicon include General Inquirer, Linguistic Enquiry and word count (LIWC), SentiWordNet etc. This method is easy to implement but does not handles the complexity of natural language well. For example- "I didn't like the movie's cast." with the lexical approach the keyword "like" will evaluated as a positive sentiment in contrast with the actual meaning of the sentiment. Advantages of this method includes its nature for being domain independent.

- MACHINE LEARNING APPROACH: Machine learning approach is where we apply certain algorithms to train our machine for making out meaning over the set of any data and help in prediction of nature of data that may further appear in future. The main idea behind machine learning approach to find a generalized idea about what the data describes by identifying patterns. In natural language it's not as easy to identify frequent patterns and hence this method's performance may vary. In some situations it may also misinterpret the data especially in the cases where unseen data is encountered.
- Machine learning can be further categorized into two categories:
- SUPERVISED LEARNING: In supervised learning we make use of labels to train our classifier that means we require two sets, training set and test set. Training set uses the classifier to differentiate the characteristics among the documents, while the test set checks the performance. First data for training set it collected and then classifier is trained accordingly with the help of the technique chosen .The predominantly used techniques to train classifier include Naive Bayes classifier, Support vector classifier, Maximum entropy model etc. The main disadvantage with supervised learning is that it requires large amount of annotated data set.
- UNSUPERVISED LEARNING: - The problems of domain dependency, multi-language applicability and human annotation requirement can be alleviated with the help of supervised learning. As described by Turney [2002] selected two arbitrary seed words (poor and excellent) in conjunction with very large text corpus. The semantic orientation of the phrases is calculated with their association with the seed word. Taking the average of all such phrases can determine the overall sentiment of the document. Supervised learning has been found to perform better than Unsupervised learning but Unsupervised learning benefits with the less cost required for it because it does not require large set of labeled data
- HYBRID TECHNIQUES

  Many researchers have tried hand at using combination of lexicon based and learning based methods. The advantages include higher accuracy hybrid techniques get the advantages from both the sides. The lexicon based method missed in capturing the actual feeling of any expression and lexicons are not good idea to rely for extracting feeling from any text. For example "funny movie" gives positive review about movie whereas "funny taste" gives away a negative meaning. Hybrid method allows to first mine the positive and negative expressions which can be further modified and improved for better results by user.

E. A new approach - Lexicon based Sentiment Analysis Model using Aspect Catalogue

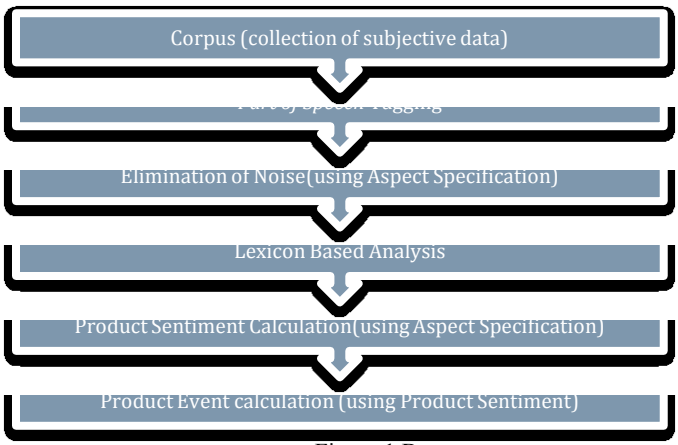

Figure 1.B

The advantages of lexicon based sentiment analysis is that it is generic the reason being the fact that the meaning that any word holds does not change as the subject changes. As stated earlier lexicon based approach is where we use dictionaries of words annotated with their semantic orientation and their polarity. The disadvantage with this is that the polarity of any word can change according to context. For example " the only problem with this city is its high cost of living" and "this phone captures high resolution images" in both the sentences the word high does not change its meaning but in the first case "high cost" is a negative sentiment while "high resolution" is a positive one. Our model is focused on alleviating this problem by first checking for the context in any given expression and then further deciding whether or not it gives positive or negative sense. The basic idea behind this model is that first all the aspect of any product have to be prerequisite determined along with their weightage and their polarity. This is a onetime process. Once this has been done we can further move on to the actual corpus for analysis. The main advantage of defining aspects in advance is that we can easily ignore about those expressions which have trivial role to play with respect to the subject or product. This can save us a lot of time and computation in determining the irrelevant expressions about the product.

Here it also needs to be noted that aspect determination is easy in some domains such as gadgets, automobiles etc. whereas it might be difficult in areas like microblogging platforms. We have as of now restricted our discussion choosing "cell phones" as our subject.

- Aspect Catalogue: - By aspects we mean features related to any subject or any product. For example for a mobile phone as a product under review its feature will include design, camera resolution, processor, screen resolution etc. It is very important to note here that we must also identify how much any aspect may matter to any user about the product. For example " voice quality of any cell phone is really shoddy but it has quite impressive design" here we can negative orientation for the aspect voice quality but positive polarity for the design but what remains completely ignored here that voice quality has more importance than design because a user would never want to compromise with that. That makes the need of aspect's weightage acknowledgement.
- Aspect weightage: - Aspect's weightage is a measure of how important any aspect is in relation to any product. With the help of aspect weightage we try to numerically treat its importance with respect to the subject. Aspect weightage can help us determine with watch intensity the negative or positive sentiments should affect the Semantic Orientation.

Sometimes positive description about trivial features can lead us to calculate wrong Semantic orientation.

- Aspect polarity :- To explain the aspect polarity let us consider the following example expressions "low cost" and "high cost" here we refer to any lexicon such as "SentiWordNet" suggest that the word "low " has a negative value ranging from 0.125 to 0.825 and the word "high" has positive as well as negative value equivalent to 0.125 and 0.25 respectively. This creates ambiguity since we haven't looked into the aspect yet. Here "low cost" must give a positive sentiment whereas "high cost" must give a negative sentiment. To check this problem we have here tried to define the polarity of aspect as well. Here the aspect polarity will help us evaluate the true meaning as well. Since for an aspect like "cost" it must be ideally low to give a positive sentiment we have defined its polarity to be negative and therefore "low cost" can be interpreted as a positive sentiment.

The following table1 illustrates aspect polarity for some features for cell.

Table 1 – Example Aspect Catalogue

Aspect	Weightage(W)	Polarity(P)
Features	0.9	positive
Looks	0.5	positive
Durability	0.75	positive
Cost	0.8	negative
Customer services	0.3	positive
Lags	0.4	negative

F. Preprocessing

The preprocessing for text required before analysis is given as follows:

- Parts of speech tagging: - parts of speech tagging will help us identify the various elements of any given sentence or expression. We have done POS tagging using NTLK (Bird, 2006).
- Elimination of Noise

"All the expressions in which there are no keywords that relate with the listed aspects in the 'Aspect Catalogue' is considered as Noise in this context"

Once the Parts of speech tagging has been done which helps us identify the nouns, adjectives, adverbs and other components of the sentence. We look for those nouns that appear in our aspect catalogue. All the other sentences and expression are disregarded. That means only those expressions which may interest the Aspect catalogues are considered for our Lexicon for Analysis.

III. LEXICON ANALYSIS

Lexicon can be either created manually or can be created automatically as well. We have made use of SentiWordNet which is a free available resource for opinion mining. It can categorize the text under the following labels:-

- Positive
- Negative
- Objective: - Any term which does not have any positive or negative meaning.

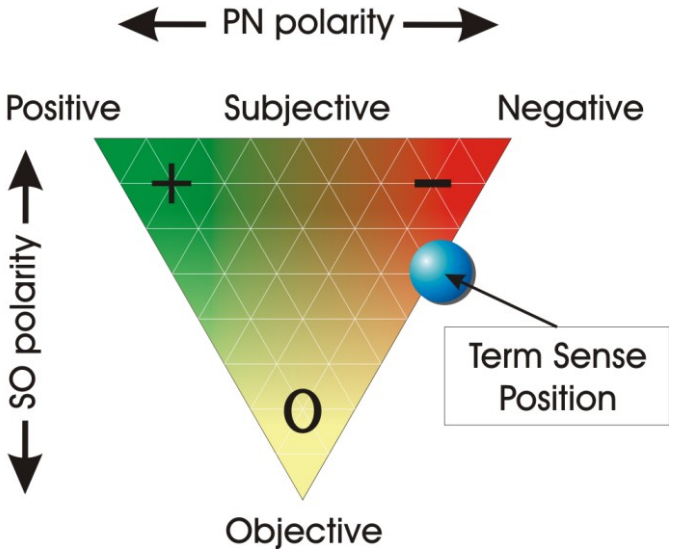

Figure 1.C

- Product Sentiment: The Product Sentiment of an opinion on a feature f states whether the opinion is positive, negative or neutral.

- Product Sentiment calculation

Let W be the weightage of any aspect where 0<=W<1

And P be the polarity of the adjective relating to the aspect where -1< P<=1

Then we calculate the Semantic orientation SO to be

$$PS = | (\sum (W*P)/\sum W) |$$

Where PS is Product Sentiment
W is Weightage
P is Polarity

IV. IMPLEMENTATION AND RESULT

A. *Aspect catalogue weightage selection*

The criteria for calculation of weightage of aspect in our model is user defined. We have chosen mobile phones review as our domain for implementation of this method and we have manually created an aspect catalogue. The weightage of features have been chosen on the basis frequency of appearance of feature in the reviews of cell phone. We have taken dataset of 500 reviews for cell phones and calculated the frequency of appearance of the features listed below. The features having the most frequency have been given more weightage considering the fact that they are the most talked about features of domain cell phones.

Table 2 manually created Aspect catalogue for "cell phone" being product for analysis

Serial no	Feature	Frequency
1	Display	467 =0.934
2	Camera	389 = 0.778
3	Sensors	73 =0.146
4	Battery	488= 0.976
5	Software issues	345 =0.69
6	RAM usage	70 = 0.14
7	Processor / performance	280= 0.56
8	Price/cost	480 =0.96
9	Touch Response	150 =0.3
10	Headphones	130= 0.26
11	Game rendering	300= 0.6
12	Speaker sound	170 =0.34
13	Storage	300 =0.6
14	Weight	150 =0.3
15	Build	380 = 0.76
16	Speed	480 =0.96
17	Service center	369 =0.738
18	Design	450 =0.9
19	Glass	370= 0.74
20	User Interface	150 =0.3
21	Color	245 =0.49
22	Sim support	175 =0.35

*467=0.934 means out of 500 reviews that particular aspect was mentioned 467 times hence giving a 93.4% frequency of occurrence.

Table 3 Calculation and Result for proposed method's accuracy

S. No	Dataset (500 review each data set)	accuracy
1	Dataset 1	60%
2	Dataset2	48%

B. *Basic Guidelines for Sentiment Analysis Model (SAM)*

The above discussed analysis model 'analyzes' and provides the user with the product that we call 'Product Sentiment' along

with the highlighted keywords that have the highest frequency. But using a supervisor or analyst who reports these Product Sentiments, beats the entire purpose of using a Sentiment Analysis Model which is to eliminate human resource consumption to provide subjective of a text, review or even a tweet. So, in order to utilize the entire might of the Sentiment Analysis Model, we need to integrate it with the core model (like Tweet based suggestions, Search Engine Optimization etc.) and install an 'Integration Algorithm' which decides how and when to use the 'Product Sentiment'.

Integration Algorithm

i. Design of the 'Aspect Catalogue' according to the core model
ii. Design a list of possible 'Product Sentiments' and 'Product Events' and assume

$$\text{Domain} - \{PSX \mid 0<x<n\}$$
$$\text{Range} - \{PEx \mid 0<x<n\}$$

Where PS – Product Sentiments
and PE – Product Events
iii. Design a relation between the Sentiments and Events as the relation R: PSX -> PEx varies according to the core model.
Note – The above relation R can be many-to-one, one-to-one, one-to-many. It can't be many-to-many as it will only confuse the core model as to what to assume as output.
iv. Decide database and executive access for each Product Event.

Even after this integration process and a heavy success rate of the above algorithm, some guidelines should be followed in order to ensure a smooth end user experience (or any level user, for that matter).

Some general guidelines that are to be followed are given below
Functioning of the Sentiment Analysis Model (SAM) should be abstract i.e. hidden from the user.
Reason for implementation – Even though the working of the SAM should be visible to some level of user (like DBA), the majority of users in social data mining environment are non-technical end users and they don't need to be aware about the SAM's existence and functioning , instead they only need to be impressed by it.

Product Events of the analysis should be non-harassing.
Reason for implementation – Again, if the end user is continuously harassed by the Product Events they will have an adverse user experience and hence it will lower the aesthetic feel of your core model. This guideline wholly depends on the Integration Algorithm.

SAM should not affect or disturb the core functionality of the model.
Reason for implementation – SAM is just a data mining model that your system can easily function without, hence the former must never affect the functioning of the latter. Unless the entire core model depends on the SAM (like Tweet based suggestions).

SAM access should be monitored and strictly implemented.
Reason for implementation – If the end user has read and write access to the data dictionary, he can easily manipulate interpretation of words (like change of word excellent from positive indexing to negative indexing), the product sentiment and hence the product events will be metaphysically wrong. Hence these read and write access to the SAM should be monitored and given only to the highest level of personnel (like the DBA, Meta-Analyst etc.). Hence these guidelines are to be followed pre-integration of SAM into your core model.

C. Applications of SAM (Post Integration)

Now that we have seen how to integrate the Sentiment Analysis Model into the core model, we are still unsure what these core models are and can be. Hence some of these core models are listed below

- Tweet based suggestions: Twitter is one of the major expressive social medium in today's connected yet distributed network. If we were to analyze these tweets and product sentiment of the particular tweet will tell us the mood of that particular twitter user.
- Message based suggestions: In today's revolutionary age of internet messaging, social media mediums like Facebook Messenger, WhatsApp etc. we can easily analyze the Product Sentiment of each message sent or received, and hence the mood of the user and suggest or disturb his machine (like smartphone , PC etc.) accordingly.
- SEO on Search Engine Side: Matching the input keywords of the end user with meta-description and meta-titles of a humongous amount of website is a very daunting task by itself, let alone adding an SAM to it. But if the hardware is upgraded to accommodate the SAM into the deployed code, the result is more end user satisfaction as the product sentiment of the input keywords can be easily matched with the product sentiment of the meta-descriptions of the websites.
- SEO on the Web developer's side: Use of SAM model to analyze the requirements of user i.e. a user of your demographic and creating a description with best possible

product sentiment to be matched with the product sentiment of the end user's input keywords.

- Advertisement Model: The SAM can be used to generate intents of the user, 'what he/she wants' , 'what he/she may need' , 'what he/she doesn't want' enabling the Advertisement model to behave according to that intent.

- Basic Subject Analysis
  The most basic use of sentiment analysis can be done on pure subjective matter such as reviews, answers as well as academic paper such as yours truly.
- The uses or application of SAM integration are countless as the number of 'core data models' are not limited to the five applications given above.

V. CONCLUSION

The main conclusion that we have derived is that adding aspect weightage to lexicon based techniques can help attain more accurate ratings. Lexicon based methods are more robust as compared to other methods as well as they are domain independent too. But we found that adding Aspect weightage adds the benefits from a domain dependent models, hence it can be conclusively said that this additional feature gives better results in the case of consumer product domain. The reason being simply the fact that consumer product review contains mostly factual information and hence it's easy to identify aspects or features related to any product in the corpuses.

We also found that substantial amount of expressions got discarded as irrelevant material with respect to the product and hence this aspect catalogue fails us in the cases where any new keyword is used to describe the product's aspect. Another disadvantage is that it requires bulky human efforts to build Aspect Catalogue, which is not a viable idea.

On top of all that we also learn how we can use the output of this Sentiment Analysis Model to generate Product Events when this Sentiment Analysis Model is coupled with a core model.

We also discussed some of these core models and how a Sentiment Analysis Model will increase their appeal and service quality towards the end user.

VI. REFERENCES

[1] Prabhu Palanisamy, Vineet Yadav and Harsha Elchuri Serendio: Simple and Practical Lexicon based approach to Sentiment Analysis.

[2] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processingVolume10, 79–86, Association for Computational Linguistics.

[3] Peter Turney. 2002. Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, 417– 424, Association for Computational Linguistics.

[4] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andr´es Montoyo. 2010. A survey on the role of negation in sentiment analysis. Proceedings of the workshop on negation and speculation innaturallanguageprocessing60–68,Associationfor Computational Linguistics.

[5] Steven Bird. 2006. NLTK: the natural language toolkit. Proceedings of the COLING/ACL on Interactive presentation sessions 69–72, Association for Computational Linguistics.