Sentiment Analysis for Helpful Reviews Prediction

Mushtaq Ahmad

May 14, 2020

Sentiment Analysis for Helpful Reviews Prediction

Author: Mushtaq Ahmad
Email: webeng.mushtaq@gmail.com
Riphah International University , Pakistan

ABSTRACT

Nowadays, every purchase we plan can be alleviated by the advice of those that tried in the past the given product. As more and more reviews are available, it would be practical to filter the relevant reviews not only to speed up the decision process but also to improve it. Gathering only the helpful reviews would reduce information processing time and save effort. To develop this functionality we need reliable prediction algorithms to classify and predict new reviews as helpful or not, even if the review has not been voted yet. In this paper, we propose a new approach which predicts reviews helpfulness based on sentiment analysis. Our approach focused on sentiment features such as the degree of positivity and the degree of negativity, in addition to the simplistic counts computed directly from reviews. It also extracts emotions dimension by means of emotion lexicon. We proposed a solution to internally construct an emotion lexicon in order to overcome challenges of invented terms, domain dependency, and spelling mistakes. We applied the proposed approach to Facebook pages of six medical products. We obtain a prediction accuracy of 97.95% through SVM algorithm. We found that sentiment degree and sadness emotion are the most decisive sentiment features to predict review helpfulness. The word count and frequencies are important as they reflect the richness and the seriousness of the review, but sentiment and emotions are more decisive as they engage and influence users.

Key words: Emotions, Facebook pages, On-line customer reviews, Reviews helpfulness prediction, Sentiments, Social media.

1. INTRODUCTION

With the advent of the Internet, people actively express their opinions about products in social media, blogs, and website comments. Online consumer reviews keep playing an increasingly important role in the decision process of buying products. Nowadays, before buying a product or a service, the costumers consult the reviews to learn from others' experience. Sometimes they have a specific question so they read carefully the available reviews trying to find a relevant response. For them, simplistic statistics like stars rating are not enough to make their decisions. Likewise, the producers are interested in reviews to keep their costumers satisfied and to evaluate their products regarding the new needs and trends. However, the available reviews are not all useful to take the decision. Gathering only the most helpful reviews would reduce information processing time and save efforts.

There is a crucial need to provide reliable prediction algorithms to classify new reviews which have not been voted but are potentially helpful. According to Park and Lee [1], users are relying on the review quality more than the quantity. They first rely on the volume of reviews a product receives to evaluate its popularity. That is to say, if the product is enough tested by others or not. Then, it gradually gets more relevant to users to read others' feedbacks to learn from their experiences and to decide. However, while more and more reviews are available online, even if people try to read the maximum, they read fewer reviews on average. For this reason, according to Malhotra et al. in [2], it is very important for retailers to provide the best reviews about the product to avoid information overload. In this context, there is a great synergy between best reviews and helpful reviews.
It concerns the most helpful reviews present on social media.

Several previous works explored the importance of the reviews to modelize helpfulness. In [3]-[4]-[5], authors studied the impact of reviews content from different perspectives. However, they merely focused on statistics computed directly from the text. Regardless of the final purpose, recent works explore emotions as an interesting way to classify documents [6]-[7]. Thus, in [8], Lionel and Pearl tried to provide a better alternative based on emotions extraction and analysis. They admit that emotions are powerful tools for communication as they evoke the feelings of others and engage their responses. Then, they justified that emotions drive people's action and regulate their decision process. Lastly, a very recent work [9] studies the role of emotions for the perceived usefulness in an online customer review. It shows how other customers are affected by customer feedback and emotions.

In this paper, we propose a new approach which predicts reviews helpfulness based on sentiment analysis. Unlike many previous works, we did not rely only on simplistic statistics computed directly from the reviews. We explored the feasibility of sentiment analysis tasks namely intensity classification and emotions extraction, to perform the prediction of helpful reviews. We first collected data about medical products from Facebook pages. Then, we automatically labeled the reviews as helpful or not based on the users' interactions and engagement such as likes and responses. For the intensity classification of sentiments, we used SenticNet [10] which is a concept-level sentiment analysis framework largely recommended in the literature [11]-[12]. For the emotions extraction, we proposed a new method based on reactions that Facebook reviews receive (like, love, haha, wow, sad, and angry). Based on those reactions we create an emotion lexicon which contains emotional terms and its scores. So if a review has not any reactions we can conclude its emotions based on the constructed lexicon. The emotional lexicon and labeled dataset will be publicly available. Through the conducted experiments, we succeed to show that both sentimental and emotional features improve helpful reviews prediction performance.

2. RELATED WORKS

The ultimate tasks of sentiment analysis are polarity classification, intensity classification, and emotion identification [13]. Polarity classification aims to classify the sentiment polarity as positive, negative, or neutral. Intensity classification seeks to identify the polarity degree to decide if the sentiment is very positive, positive, fair, negative, or very negative [14]. Further, emotion identification attempts to identify the specific emotion behind the sentiment such as sadness, hanger, and fear [15]. Sentiment analysis through machine learning and natural language processing techniques has become a popular method to extract features from the User-Generated-Content (UGC) [22]. Existing studies have shown a relationship between sentiment analysis and review helpfulness [3]-[4]-[16]. In [4], Mudambi and Schuff collected 1587 reviews of six products from Amazon.com and found that the intensity of polarity can affect review helpfulness. Hu et al. [16] empirically compared the relationships among review rating, sentiment, and product sales by analyzing book reviews at Amazon.com. The results showed that sentiment features have a strong relationship with product sales. The most helpful reviews have an important influence on sales.

In the same context, Hwang et al. [17] investigated the effects of different features on hotel reviews helpfulness prediction. A total of 3124 hotel reviews were collected from TripAavisor.com and manually labeled into helpful and not helpful. Three kinds of features were retrieved from the review text namely: TF-IDF, topic-model-based Latent Dirichlet Allocation (LDA), and semantic-based LDA features. The last two methods utilize LDA technique to generate a set of topics from the set of reviews. Then, each topic is associated with a multinomial distribution over words. The authors considered also sentiment features in their study. However, topic-model-based LDA approach showed the best classification technique due to its relatively higher recall rate and F1 with the use of fewer content features.

Instead of exploring more sentiment features as well as the emotional ones, Zhu et al. [18] explored more statistical information. They investigated the relationship between reviewer credibility and review helpfulness and the moderation effects of hotel price and review rating extremity. They collected a total of 16,265 hotels reviews from Yelp.com and automatically labeled them based on users' votes (stars on reviews). However, in this case, reviews written by opinion leaders would receive more votes (if the reviewers gave these reviews moderate star ratings). So that reviews written spontaneously by particular users are not considered which would lead to a loss of important information.

Although the above studies have identified several predictors for review helpfulness in different domains (hotels, products, and sales), they did not consider the link between the different kinds of features. Exploring sentimental and emotional features together with statistical features such as TF-IDF would lead to more comprehensible and interpretable results. The relationship between statistical features such as review length and sentiment features such as the degree of positivity should be investigated in order to provide a more accurate helpfulness prediction.

3. RESEARCH METHOD

Figure 1 illustrates the research process, which can be divided into five main steps: Facebook pages extraction, Reactions analysis in order to construct an emotional lexicon, Reviews processing and automatic labeling based on the total reactions, Features extraction, and prediction model construction.

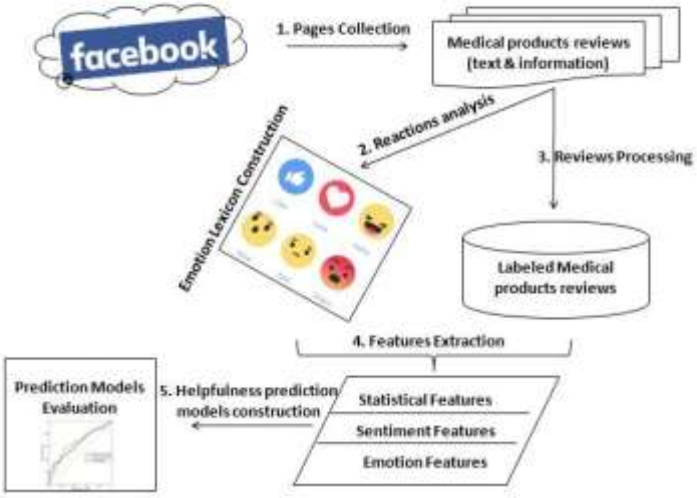

facebook
1. Pages Collection
Medical products reviews (text & information)
2. Reactions analysis
3. Reviews Processing
Emotion Lexicon Construction
Labeled Medical products reviews
4. Features Extraction
Prediction Models Evaluation
5. Helpfulness prediction models construction
Statistical Features
Sentiment Features
Emotion Features

Figure 1: Research process.

3.1 Data collection

Social media platforms have encouraged user-generatedcontent production. Thus, it has been a growing interest in big social data analysis within both academic research and business world. Social media platforms represent an innovative tool to achieve insights into behavioral patterns of users. Cambria et al. in [19] studied the impact of sentiment analysis on social media to extract useful information from unstructured data to evaluate consumer products and financial services.

In our research, we choose Facebook as a source to provide a labeled data set to predict reviews helpfulness. We collected data from Facebook using Facebook Graph API. We gathered reviews published on pages of known medical products in U.S.A namely life's DHA, Medtronic Diabetes, NeilMed, AdvoCare, Nature's Bounty, Clearblue, and Zarbee's Naturals. From six verified Facebook pages, we extracted more than 3k post, when each post provides thousands of reviews. We did not only extract the review text, but also the available metrics such as likes count, responses texts and metrics, and timestamps.

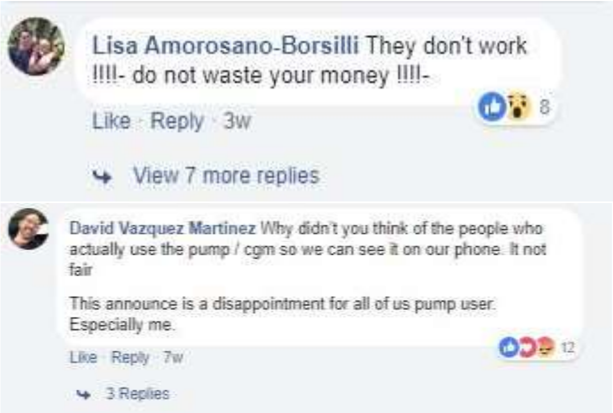
Figure 2 is a snapshot of a medical product reviews from Facebook pages which shows the information we need for this research. The information includes:

The review text: The text may contains emoticons like ☺ and☹, several punctuations (., !, ?), and misspelled words.

Reactions count: There is a set of six reactions that users use to vote to the review regarding their emotion (like, love, haha, wow, sad, and angry).

Responses: Instead of reactions, users can react by writing texts. In its turn, a response is a review which my receive reactions. So, we extracted the response count, response texts, and reactions to the response.

Timestamps: We extract the time of the review as well as the time of response. This information helped us while labeling the dataset.



Lisa Amorosano-Borsilli They don't work !!!!- do not waste your money !!!!-

David Vazquez Martinez Why didn't you think of the people who actually use the pump / cgm so we can see it on our phone. It not fair

This announce is a disappointment for all of us pump user. Especially me.

Figure 2: Example of medical products reviews from Facebook pages.

3.2 Dataset labeling

Some previous works chose to do manual labeling by the intervention of domain expert. However, the manual labeling is expensive and time-consuming. Others work conducted an automatic labeling by considering reviews rating (five stars note) as users vote. So that reviews with one or two stars are considered unhelpful and reviews with four or five stars are considered helpful. However, those proceedings are confused. First, a review with one or two reviews may be helpful but not very helpful. Second, the review may receive more votes after the data mining time.

In our research, we considered users reactions and responses as votes. Reactions and responses reflect users' engagement toward the review. So that very ancient reviews with no reactions and responses were considered unhelpful. Here, we used the timestamp to select a large set of ancient reviews and filter out reviews with users' engagement. In contrast, although their recency, there is very recent reviews with many reactions and responses. Those last are considered as helpful. Our dataset contains 10,019 reviews when 5006 labeled helpful and 5013 labeled unhelpful.

3.3 Emotion Lexicon construction

In social networks, if someone reacted to an entity (public post, review, or post), it means that the person has positive or negative feelings towards the entity in question [13]. Those feelings may be expressed explicitly through reviews or implicitly through reactions. Hence, we explore Facebook reactions to construct an emotion lexicon. This resource would allow detecting emotions in any review, even if the review did not receive any reaction until a specific time.

There are six kinds of reactions that Facebook users use to express their emotions. So, each reaction serves to learn about users' emotions and feelings. According to Liu in [14], emotions represent our subjective feelings and thoughts which arise in response to appraisals one makes for something of relevance to one's well-being. In the case of Facebook, if the user feels well, he would 'like' the review. If he feels very well, he would 'love' the review. Likewise, if he found the entity bad or very bad, he will click respectively on 'sad' or 'angry'. The user may also be surprised so he would click on 'wow', or laugh so he would click on 'haha'.

From the collected data we selected all the reviews which received any kind of reaction. After cleaning the review and deleting stop words, based on the reactions, we selected terms reflecting emotions. In this level, we got a list of emotional terms. Based on reactions count, we give a score for each term. For example, the term 'satisfied' appear in review 1 and review 2. Review 1 has 5 likes, 20 loves, 0 haha, 1 wow, 0 sad and 0 angry. Review 2 has 15 likes, 40 loves, 1 haha, 12 wow, 0 sad, and 0 angry. So, the term 'satisfied' has 20 likes, 60 loves, 1 haha, 13 wow, 0 sad, and 0 angry. The normalization was done by the means of the sum of all reaction. In this example, the term satisfied has 60/94 like (0.64). The constructed emotional lexicon contains 23, 6899 unique terms with corresponding like, love, laugh, surprising, sadness and angriness ratios.

Using external resources to analyze extracted data from social media (including Facebook) would be challenging due to the non-dictionary words, colloquial terms, spelling mistakes, and domain dependency. It may happen that misspelled word is often used in a specific domain. Moreover, a product name written using upper case letters or redounding letters inform us about the emotion of the reviewer. Our solution overcomes those challenges as it is an internally constructed resource. Our lexicon contains frequent misspelled words (luvvvv instead of love), terms related to the domain (advocaaaaare), and invented terms (lol, zzz…).

3.4 Features extraction

Predictive features are considered as the main elements affecting the performance of supervised classification. In our research, we characterize review helpfulness through three kinds of features. In total we designed 17 features: 7 statistical features, 3 sentiment features, and 7 emotional features. Statistical features computed directly from the text such as review length in term of words and sentences would help the prediction but are very simplistic to be very accurate.

Table 1: Features categories and details

Category	Features	Description
Statistical	WC	Word Count: Length of review in term of words.
	SC	Sentence Count: Length of review in term of words.
	WSR	Word per Sentence Ratio.
	SMC	Spelling Mistakes Count.
	IC	Interrogation Count.
	EC	Exclamation Count.
	CC	Comma Count.
Sentimental	SD	Sentiment Degree.
	PER	Positive Emoticons Ratio☺.
	NER	Negative Emoticons Ratio☹.
Emotional	EMR	Emotion Ratio: Emotional word count.
	LKR	Like Ratio.
	LVR	Love Ratio.
	LGR	Laugh Ratio.
	SPR	Surprised Ratio.
	AGR	Angry Ratio.
	SDR	Sad Ratio.

Emotions and sentiments exploration is crucial to build a more accurate predictor. According to Garcia and Schweitzer in [20], human beings are empathetic creatures that perceive emotions as information comparable to factual data, which makes emotions a valuable additional feature set. Hence, in our research, we worked on three features categories illustrated in Table 1.

The statistical features are directly computed from the review text. Through WC, SC, and WSR we evaluate the review length from different angles. SMC serves to evaluate the review in term of spelling mistakes, as many mistakes would make the review difficult to understand. IC and EC feature respectively investigate if interrogation and exclamation punctuation provide important information. While the presence of comma is related to detailing things, reviews containing many commas would provide much information about the product. Therefore, the CC feature would help to model helpful reviews.

Sentiment features are mainly based on positivity and negativity degree, and emoticons presence. To compute the sentiment degree of each review, we used SenticNet Framework which allows classifying the polarity with intensity by attributing a degree between -1 and 1. While helpful reviews may be positive as negative, considering merely the polarity (+ or -) rather than the degree would not provide special information. Therefore, SD feature values represent the sentiment degree of each review. Besides the SD, we design two others sentimental features based on emoticons. PER and NER are respectively the ratio of positive emoticons and the ratio of negative emotions. It may happen that a review contains both positive and negative emoticons at the same time.

Emotional features detail more the sentiment dimension. A positive sentiment may reflect like, love, or both of them with different degree. So the two reactions ‘like’ and ‘love’ allow as extracting these two emotional dimensions. Likewise, negative sentiment may reflect sadness, angriness, or both of them with different degree. We extracted sadness and angriness dimensions based on ‘sad’ and ‘angry’ reactions. Furthermore, the user may be surprised, or laugh on something. The two reactions ‘wow’ and ‘haha’ allowed us to extract these last emotions. Based on the constructed emotion lexicon (section 3.3), we extract the first emotion feature EMR. This feature represents the ratios of the emotional terms present in the review. Then, based on the same lexicon, we extracted LKR, LVR, LGR, SPR, AGR, and SDR features. They respectively represent the ratio of like, love, surprise, laugh, angriness and sadness emotions present in the review. One review may contain one or more emotions with different ratio.

4. EXPERIMENTAL EVALUATION

In this section, we compare the performance of several supervised classification algorithms to select the best one. Then, we study the feature importance in order to show how the designed features characterize reviews helpfulness. In [23], Hari Krishna Kanagala et al. provide a useful review on classification techniques in data mining. In all our experiments, we used machine learning algorithms from scikit-learn package [21].

4.1 Learning quality

Unlike many previous works, our labeled dataset do not seek from class imbalance. The helpful class examples (5006) and the unhelpful class examples (5013) are almost equal. So we do not need to deal with class imbalance issue. This advantage refers to large amount of data that Facebook pages provide. Since our dataset is balanced, we performed the learning phase with confidence to obtain a model that accurately fits our objective. We experiment several supervised classification algorithms (SVM, Random Forest, Naïve Bayes, and Neural Network). Then, we selected the best algorithm in terms of accuracy (ACC), F-measure (F1), and area under the curve (AUC). Classifier performances are reported in table 2. In term of ACC, F1, and AUC, the SVM algorithm achieves the best performance followed by Random Forest and Naïve Bayes. The Neural Network showed the worst performance in terms of ACC, F1, and AUC. For more details about classifiers performance, we present in table 3 the confusion matrix of each algorithm (true positive TP, false positive FP, true negative TN, and false negative FN).

Table 2: Performance comparison of various classifiers.

Classifier	TP	FP	TN	FN
Random Forest	92.83	07.17	99.56	00.44
SVM	**99.88**	**00.12**	**96.30**	**03.70**
Neural Network	79.86	20.14	29.72	70.28
Naïve Bayes	89.55	10.45	81.31	18.69

Upon visual inspection, we plot in the same graph the ROC (AUC) curve of each classifier. The abscissa axe represents the false positive rate and the ordinate axe represents the true positive rate. As illustrated in Figure 2, the curve of SVM is closer than others classifiers curves to the upper-left corner of the ROC space. This means that SVM achieve the best trade-off between sensitivity (true positive rate) and specificity (false positive rate). It shows the best performance to correctly predict the helpful class with minimal false positive (classifying unhelpful reviews as helpful). So, SVM is the best classifier algorithm to predict helpful reviews.

Table 3: Confusion matrix

Classifier	ACC	F1	AUC
Random Forest	96.20	96.32	96.19
SVM	**97.95**	**97.92**	**97.96**
Neural Network	54.78	39.68	54.79
Naïve Bayes	85.43	84.81	85.43

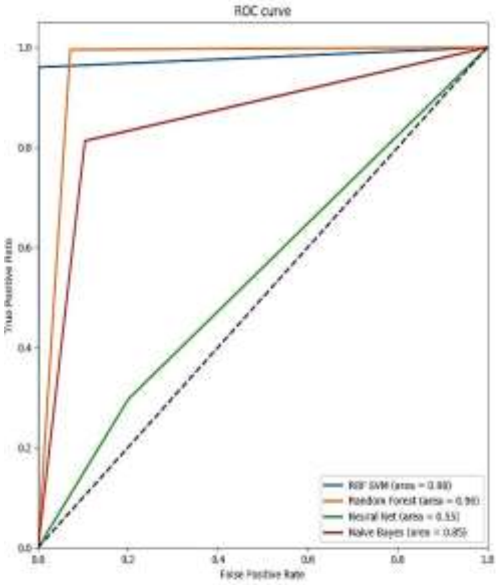


Figure 3: Performance classifiers using ROC curve.

4.2 Features importance

In this section, we aim to compare the relevance of the designed features through its prediction strength by categories. To measure the prediction strength we draw the features importance plot in Random Forest classification.

Figure 4 illustrates statistical features importance. We notice that the feature word count (WC) and the feature word per sentence ratio (WSR) are the most important features, followed by sentence count (SC), exclamation count (EC), and commas count (CC). WC and WSR allow evaluating the richness, the clarity and the consistency of the review. The more these properties are present, the more the review is helpful. SC, EC, and CC are slightly helpful. It seems that the punctuation does not give clear information about the helpfulness. People in social media tend to be more informal to use properly the different kind of punctuation. Lastly, spelling mistakes (SMC) and interrogation count (IC) are poorly decisive because people in social media are used to make mistakes. In addition, when the review contains questions, it means that the review does not really contain information. Instead, the reviewer is asking for information.

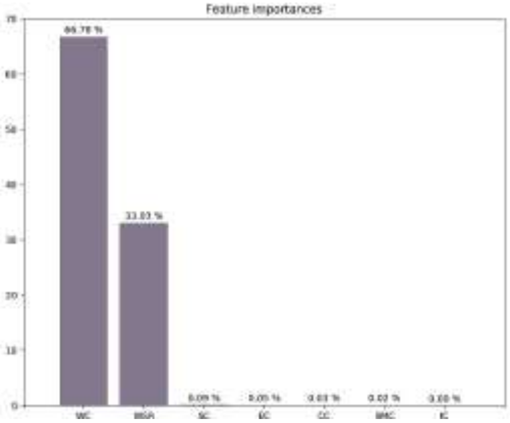


Figure 4: Statistical features importance.

Figure 5 illustrates sentiment features importance. We notice that the feature sentiment degree (SD) is highly decisive. This feature gives details about the polarity of the sentiment rather than be merely binary. In contrast, the two features positive emoticons ratio (PER) and negative emoticons ratio (NER) are very poorly decisive compared by SD. In the future, we should consider others informative emoticons rather than explored only positive and negative emoticons.

Figure 6 illustrates emotions features importance. We notice that the feature sadness ratio (SDR) is the most decisive feature among the emotional features, followed by emotions ratio (EMR) and surprise ratio (SPR). It seems that users concentrate more on negative reviews as they tend to share their bad experience more than the good ones. Hence, we find like, laugh, and love ratio features are significantly less decisive. However, although it is negative, the angry ratio features (AGR) is the less decisive one. It seems that angriness emotion make people speak shortly but vulgarly. So they did not share their experiences properly to make others benefit from it.

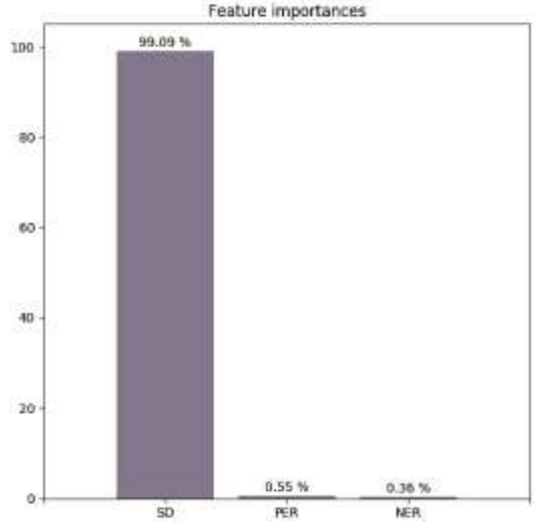


Figure 5: Sentiment features importance.

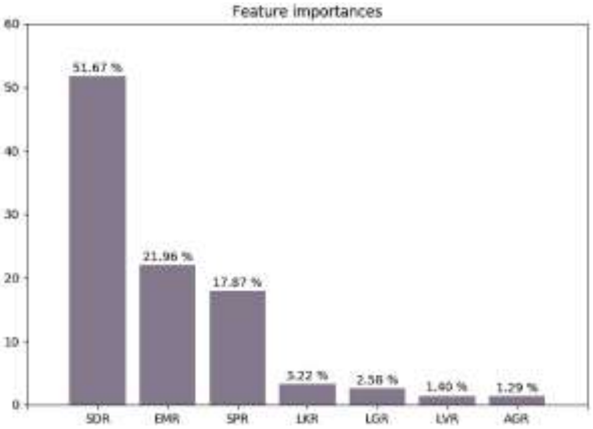


Figure 6: Emotion features importance.

5. CONCLUSION

We develop a method for predicting the helpfulness of online product reviews using sentiment and emotions. We first collected reviews from Facebook pages using Facebook graph API. Then, a large emotion lexicon is built (constructed) from the Facebook reactions since emotions are likely (known) to trigger reactions. This can be used to extract emotions from a review even if it did not receive any reaction.

Based on the different kind of users' interactions such as responses and likes, we manually labeled a relatively large dataset. For each review, we computed statistical features from the review text such as word counts, extracted sentiment features using SenticNet framework, and computed 7 emotional features based on the constructed emotion lexicon.

On the results, we discussed the performance of several classification algorithms and studied the significance of all features. Among these algorithms, SVM achieved the maximum prediction accuracy of 97.95%. We found that negative sentiment and sadness emotion are the most decisive sentiment features for predicting the helpfulness of a review. Compared to the statistical features, they are more decisive since they engage and influence users while statistical features still reflect the richness and the clarity of the review.

For future work, we will focus on improving the treatment of negative emotion as we have seen that negative sentiment and emotion are more decisive for predicting the helpfulness of reviews. In addition, we will explore more the diversity of emoticons that are used on social media to cover more emotions and enhance the created lexicon.

REFERENCES

1. D. H. Park and J. Lee. **eWOM overload and its effect on consumer behavioral intention depending on consumer involvement**, *Electronic Commerce Research and Applications*, vol. 7, no. 4, pp. 386-398, Dec 2008. https://doi.org/10.1016/j.elerap.2007.11.004
2. N. K. Malhotra. **Reflections on the information overload paradigm in consumer decision making**, *Journal of consumer research*, vol. 10, no 4, p. 436-440, Mar 1984. https://doi.org/10.1086/208982
3. Q. Cao, W. Duan, and Q. Gan. **Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach**, *Decision Support Systems*, vol. 50, no. 2, pp. 511-521, Jan 2011. https://doi.org/10.1016/j.dss.2010.11.009
4. S. M. Mudambi and D. Schuff. **Research note: What makes a helpful online review? A study of customer reviews on Amazon. Com**, *MIS quarterly*, pp. 185-200, Mar 2010. https://doi.org/10.1111/j.1083-6101.2011.01551.x
5. L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. De Ridder. **"Highly recommended!" The content characteristics and perceived usefulness of online consumer reviews**, *Journal of ComputerMediated Communication*, vol. 17, no. 1, pp. 19-38, Oct 2011.
6. S. Mohammad. **From once upon a time to happily ever after: Tracking emotions in novels and fairy tales**, in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, 2011, pp. 105-114.