

Building a fuzzy system for opinion classification across different domains

Matheus Cardoso

State University of Feira de Santana (UEFS)
and Federal University of Bahia (UFBA)
Salvador, Brazil
matheus.mcas@gmail.com

Angelo Loula

Matheus Giovanni Pires

State University of Feira de Santana (UEFS)
Feira de Santana, Brazil
{angelocl, mgpires}@ecomp.uefs.br

Abstract

Opinions are central in almost all human activities, because they are a relevant influence on people's behavior. The internet and the web have created mechanisms that made possible for people to share their opinions and for other people and organizations to find out more about opinions and experiences from individuals and help in decision making. Still, opinions involve sentiments that are vague and inaccurate textual descriptions. Hence, due to data's nature, Fuzzy Logic can be a promising approach. This paper proposes a fuzzy system to perform opinion classification across different domains. Almost 70 features were extracted from documents and multiple feature selection algorithms were applied to select the most fitted features to classify documents. Over the selected features, the Wang-Mendel (WM) method was used to generate fuzzy rules and classify documents. The WM fuzzy system based achieved 71,25% of accuracy in a 10-fold cross-validation.

1 Introduction

Opinions are central in human lives. In almost all people's daily tasks, they ask or seek other people's opinions to help them make decisions, such as what movie to watch, what car, book or notebook, for instance, to buy or to know what are the political standpoint of their neighborhood about a certain issue. The internet and the web have created mechanisms that made possible for people to share their opinions and also organizations to find out more about opinions and experiences from other individuals, most of them unknown persons. This mechanisms, over the time, have created a huge amount of opinative sources, hard for a person to process by itself. Hence, an automated opinion mining system is required, one that could identify and extract opinions from text would be an enormous help to someone sifting through the vast amounts of news and web data [Wilson *et al.*, 2004].

Opinion mining is the process that seeks to predict the overall sentiment orientation conveyed in a piece of text such as a user review of a movie or product, blog post or editorial [Ohana *et al.*, 2011]. Attached to opinions, there are sentiments. Sentiments are intrinsically subjective and to identify

them in phrases and documents we have to deal with vague and imprecise terms, such as "good", "very nice", "bad", among others. Due to the nature of this data, Fuzzy Logic [Zadeh, 1965] can be a promising approach to deal with this.

Given the importance of opinions in human lives, the commercial and political relevance, the huge amount of generated data that has to be automatically handled, besides the vague and imprecise nature of the data, this paper aims to propose and evaluate an automated fuzzy opinion mining system to classify the overall sentiment orientation of a document's text. Our proposal differs from others because it generates fuzzy rules based on most fitted features among almost 70 features that were extracted from documents, introducing the use of the Wang-Mendel method [Wang and Mendel, 1992]. We apply those rules to perform opinion classification across different domains.

The next section writes about related works, describing previous works on opinion mining and applications of fuzzy logic. The following section outlines the opinion mining process, specifying all stages involved in opinion mining workflow. Results from our approach are shown and discussed next. The last section concludes this paper pointing out our contributions and some future improvements to this research.

2 Related works

The research in opinion mining began with subjectivity detection, dating back to the late 1990s, with [Wiebe, 1990; 1994]. This task involves separating non-opinionated, neutral and objective sentences from subjective sentences carrying heavy sentiments. Following the years, starting at 2000s the overall research focus has shifted to divide the language units into three categories: negative, positive and neutral. From there many works on this task, also known as sentiment analysis or sentiment classification, among other naming, has arrived.

One of the first research studies on unsupervised opinion mining was [Turney, 2002]. Similar to the task of classifying documents as positive or negative, [Turney, 2002] proposed to classify reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. He got a average of 74% accuracy across domains.

On the other hand, [Pang *et al.*, 2002] was one of the first to propose using classic machine learning techniques in opinion mining. Comparing the performance between Naive Bayes, Maximum Entropy and Support Vector Machine (SVM), this work showed that such techniques produces high accuracy levels, achieving 82.9% of accuracy using only isolated words (called unigrams) with SVM. It showed as well that supervised techniques shows better results than unsupervised approaches. However, they are domain dependent, producing even poorest results in other kind of data, demanding another training round of the classifier, increasing cost and time to classify documents.

Related to our work, there are the works from [Wilson *et al.*, 2005], [Taboada *et al.*, 2008] and [Ohana and Tierney, 2009] that use a wide range of document's features. These features range from the count of adjectives, adverbs in phrase or whole document, tuples of words (called bigrams if two words; trigrams if three words), such as adverbs and adjectives, to the sum of polarities and many others features. [Taboada *et al.*, 2008] and [Ohana and Tierney, 2009] use a semantic lexicon, Sentiwordnet [Esuli and Sebastiani, 2006], to assign numeric values to word's semantic orientation. In classifying documents as negative or positive, the results obtained were 65,7% of accuracy in [Wilson *et al.*, 2005], 80.6% in [Taboada *et al.*, 2008], and 69.35% in [Ohana and Tierney, 2009].

Although it has been shown that Fuzzy Logic is suitable to handle imprecise and vague data [Zadeh, 1996; Wang, 2003], we found only a few works applying fuzzy concepts to opinion mining, such as fuzzy sets or fuzzy inference systems. One of the few papers found was [Nadali *et al.*, 2010]. It proposes a fuzzy logic model to perform semantic classifications of customers review into five classes: very weak, weak, moderate, very strong and strong. Also introduces a methodology that implies use of a fuzzy inference system, fuzzy sets that models the five classes and manually created IF-THEN rules. However, the paper did not describe results or further discussion.

Another paper was [Ballhysa and Asilkan, 2012] that proposes a fuzzy approach for discovering the underlying opinion in entries in blogs, determining the overall polarity. The authors presented fuzzy concepts such as fuzzy sets and fuzzy sets operations. They proposed a set of fuzzy measures (from counting manually chosen keywords) and a single fuzzy aggregation of these measures, but a fuzzy inference system is not used. However, the proposed measures seem to actually correspond crisp value, so there is no actual application of fuzzy logic. Moreover, there is only a superficial description of results, obtained on their own dataset with no comparison with other works.

This paper differs from previous work on applying fuzzy systems for opinion mining. We model fuzzy variables and build a fuzzy inference system based on document features. We run our tests in datasets already used in previous works, allowing direct comparison. Besides, we propose a feature extraction and selection stage, where we extract a great number of features from documents, based on previous works and extended with our own features, and perform feature selection based on different algorithms. The next section presents the

opinion mining process that we used, describing each stage and the relevant techniques used on them.

3 The opinion mining process

Our opinion mining process is composed by five stages: domain definition, preprocessing and transformation, feature extraction and selection, classification and evaluation. In the first stage it is defined what kind of data will be handled by the system and what datasets will be used. We picked up the widely used Cornell Movie Review Data 2.0 [Pang and Lee, 2004] and a mixed dataset containing Amazon products [Wang *et al.*, 2011], such as camera, mobile phone, TV, laptop, tablet, among others to evaluate our cross domain proposal.

3.1 Preprocessing and Transformation

In the preprocessing stage, data filtering takes place and a document representation model is built. There are three basic levels of document analysis: document, sentence and entities and its aspects [Liu, 2012]. The first level focus on classify opinions as positive or negative from the whole document perspective. The second one seeks to classify opinions of each sentence in a document and the last level looks to classify opinions targeted to aspects of the found entities. We chose the document level analysis [Turney, 2002; Pang *et al.*, 2002; Pang and Lee, 2004; Taboada *et al.*, 2008].

As a first step, we remove all the sentences in a document that has modal words on it, such as "would", "could", among others. Modals indicates that the words appearing in a sentence might not be reliable for the purposes of sentiment analysis [Taboada *et al.*, 2011]. Next, all words in each document are tagged with its grammatical class using a POS (Part of Speech) tagger [Brill, 1995].

The document model in our approach is the popular bag-of-words model in which a document is represented as a vector, whose entries correspond to individual terms of a vocabulary [Moraes *et al.*, 2012]. These terms are called generically as n-grams. They can be unigrams (only one word), bigrams (two words) and trigrams (three words). For each document, one n-gram vector leaves to the next step of the process.

We defined 7 types of n-grams: adjectives, adverbs, verbs as unigrams ; adverbs with adjectives (e.g. very good) , adverbs with verbs (e.g. truly recommend), adverbs with adverbs as bigrams and one type of trigram, the combination of two adverbs with one adjective (e.g. not very nice) [Pang *et al.*, 2002; Turney, 2002; Taboada *et al.*, 2008; Karamibekr and Ghorbani, 2012].

We also look for special types of bigrams and trigrams: the negated n-grams (e.g. not bad, nothing special). This technique is called negation detection and by itself it is a entire line of research, going beyond this work scope, but we use a simple version from [Taboada *et al.*, 2011].

At this point stage, each document was transformed into a n-gram bag-of-words vector. Each n-gram is now associated with a numeric value, an opinion polarity degree, using an opinion lexicon. Opinion lexicons are resources that associate words with sentiment orientation [Ohana and Tierney, 2009]. Hence, we decided to use a automatically built opinion lexicon, the Sentiwordnet [Baccianella *et al.*, 2010].

SentiWordNet (SWN) is a lexical resource explicitly devised for supporting sentiment classification. SWN provides positive, negative and objective scores (ranging from 0 to 1) for each sense of English words. Since words can have multiple senses, we apply the approach proposed by [Guerini *et al.*, 2013], called prior priorities, to derive positive and negative polarity for words.

To determine polarity degrees for bigrams and trigrams, we consider adverbs as modifiers, subdivided into amplifiers (e.g. very) and downtoners (e.g. slightly) to increase or decrease adjective (unigram) values, respectively [Quirk *et al.*, 1985]. Downtoners and amplifiers have sub-levels, each of them has a modifier value associated, such as -0.5 to "lowest" downtoners and 0.25 to "high" amplifiers, among others sub-levels. The final score s for a bigram is defined by $s(\text{bigram}) = s(\text{unigram}) + s(\text{unigram}) \cdot s(\text{modifier})$ and score s for trigram by $s(\text{trigram}) = s(\text{bigram}) + s(\text{bigram}) \cdot s(\text{modifier})$.

The special case among bigrams and trigrams are the negated ones. For these, instead of use modifiers, we apply a similar approach made by [Taboada *et al.*, 2011], shifting the n-gram polarity to the opposite sign by a fixed amount (0.5, empirically defined). [Taboada *et al.*, 2011] has shown as well that shift polarity is better than just invert the n-gram polarity sign.

Other technique was the attenuation by n-gram frequency, in which a term polarity is decreased by the number of times that it appears in the document. The n th appearance of a word in text will have the new score s' defined by $s'(\text{word}) = s(\text{word})/n$. The repetition of an adjective, for instance, suggests that the writer lacks additional substantive commentary, and is simply using a generic positive word [Taboada *et al.*, 2011]. Also we have used a bias compensation to negative term polarities. Lexicon-based sentiment classifiers generally show a positive bias [Alistair and Diana, 2005], likely the result of a universal human tendency to favor positive language [Boucher and Osgood, 1969]. So, we increased the final n-gram degree of any negative expression (after other modifiers have applied) by a fixed amount (currently 50%). In the end of this stage, we have a vector of n-grams associated with polarity degrees for each document of the dataset.

3.2 Feature extraction

In this step, we extract document features from the previous numerical n-grams vectors to intent to be domain independent. We decided this approach, because is effective as it takes reviews, checks documents features and decides its semantic orientation considering only its characteristics, instead of its specific contents. The features we use are not specific to a domain, and should be easily applicable to other domains [Pang *et al.*, 2002]. Also, this reduces features dimensionality, since the resultant features vector is significantly smaller than a regular bag-of-words vector.

On the other hand, corpus-based machine learning methods applied to opining mining are able to obtain high accuracy rates, up to 95%, feeding word vectors directly to classifiers, which will learn from the given document corpus which words are related to positive and negative contexts. However, in order to reach their full potential, most of these approaches

need immense annotated training datasets, huge amount of time for training and still produces poorest results across domain without full retraining.

Different studies proposed many various features to describe or discriminate documents among themselves to identify their polarities [Wilson *et al.*, 2005; Ohana and Tierney, 2009; Taboada *et al.*, 2011]. In order to capture diverse aspects from documents, we decided to extract a great number of features, so we used features presented in these works and derived many others, obtaining a total of 67 features.

Three kinds of features were defined: sum, count and maximum values. Sum features involves the numerical sum of polarity degrees for different types of n-grams, such as sum of adjectives of a document, sum of adverbs, verbs, bigrams composed by adverb and adjective, sum of trigrams, among others. The count features proceeds in a similar way for different types of n-grams, counting the number of positive or negative polarity values.

The maximum values features refer to the maximum value of a given type of n-gram in a document. For instance, if the maximum absolute value among the unigrams is positive, this feature has the value 1. On the other side, if maximum value is negative, this feature has the value -1. This feature was obtained for unigrams, bigrams and trigrams.

More features were derived from the three kinds described above by applying normalization or subtraction of features. For instance, the difference between positives and negatives bigrams of a document and the normalized sum of positive adjectives are one of these derived features.

After the feature extraction step, vectors of n-grams and polarity values are replaced by feature vectors. Each document in the dataset is now represented by a 67 size feature vector.

3.3 Feature selection

This stage is commonly found in opinion mining approaches. It can make classifiers more efficient/effective by reducing feature vector dimensionality, the amount of data to be analyzed as well as identifying relevant features to be considered [Moraes *et al.*, 2012]. To choose the features among the ones extracted and reduce the amount of features to be analyzed by the classifier, we used two algorithms for feature selection, the Correlation-based Feature Selection (CFS) and feature selection from C4.5 decision tree [Cintra *et al.*, 2008].

CFS evaluates subsets of features on the basis that a suitable feature subsets contain features highly correlated with the classification, yet uncorrelated to each other [Hall, 1999]. C4.5, in other hand, is an algorithm that generates a decision tree that can be used to a classification task [Quinlan, 1993]. But, to build that tree, c4.5 needs to select the best features among the provided. Hence, we also use c4.5 as our feature selection algorithm.

3.4 Classification

In the classification stage, we build a rule-based fuzzy system classifier to predict the overall sentiment orientation, as positive or negative, of each document in the dataset. Building such classifier involves creating a set of rules based on the

extracted features, modeling these features as linguistic variables with fuzzy sets, lastly, defining an inference system.

In order for fuzzy sets to appropriately model data, we first identify outliers values in features and limit the range of feature values. To do this, we used the three-sigma rule [Kazmier, 2004] to select outliers values that lie after three standard deviations from the mean of a feature, an interval where 99.73% of the values in a normal distribution stand in. Outliers values left out this range were modified to the extreme value of the accepted range.

Now, with the input range standardized for every feature, we can define the fuzzy sets [Zadeh, 1965] to model our input and output variables. We decided to use triangular fuzzy sets. The first approach was use three fuzzy sets in the input (low, medium and high) and two sets for the output (negative and positive), uniformly distributed along the feature value range. Another approach was to use only two sets in the input, removing the medium fuzzy set.

Once fuzzy variables were modeled from fuzzy sets, the next step was to build our fuzzy rule base using the Wang-Mendel Fuzzy Rule generation [Wang and Mendel, 1992]. With previously specified fuzzy sets, this fuzzy rule generation method takes each data instance in the dataset, determines pertinence degrees in all fuzzy sets and builds rules using the fuzzy sets with highest pertinence degrees, for each input-output pair.

The generated fuzzy rule base along with the specified fuzzy sets are then used by a fuzzy inference mechanism to determine document polarity class. The mechanism used were the General Fuzzy Reasoning Method (GFRM) and the Classic Fuzzy Reasoning Method (CFRM) [Cordon *et al.*, 1999].

In this classification process, each document feature vector is evaluated by all fuzzy rules and a compatibility degree is produced for each rule. The CFRM picks up the rule with the maximum compatibility degree and assigns the rule output class to document. In the other side, GFRM takes the maximum average compatibility degree between the two possible classes, positive and negative. In other words, GFRM calculates the average degree among all rules with negative and positive output and assigns to document the class from the maximum average compatibility degree.

3.5 Evaluation

In order to evaluate our opinion classification approach, we apply a 10-fold cross-validation. As measures of classification performance, accuracy, recall, precision and F1 score were chosen. Accuracy is a measure of the ratio between documents that has been classified correctly to the total number of documents being classified. Recall measures the ratio of documents correctly classified into a category to the total number of documents truly belonging to that category. This measurement indicates the ability to recall items in the category. Precision measures the ratio of the number of documents correctly classified into a category to the total number of documents classified into category. And F1 score is a measure that considers both the Precision and the Recall to compute the score. F1 is often considered as a weighted average of the precision and recall. [Chaovalit and Zhou, 2005].

4 Results

In this section we describe and discuss our experiments and its results. We aim to not only compare best classification accuracy but also discuss contexts in which the classifiers produce better or worse results.

4.1 Datasets

We performed our experiments on two datasets, as described before. Each dataset consists of 2000 reviews that were previously classified in terms of the overall orientation as being either positive or negative (1000 positive and 1000 negative reviews). For the Amazon dataset, the ground truth was obtained according to the customer 5-stars rating. Reviews with more than 3 stars were defined as being positive and reviews with less than 3 stars were labeled as being negative. Reviews with 3 stars were not included in our analysis. In the movie reviews dataset, all documents were already tagged as positive or negative.

4.2 Design of experiments

We focus on comparing GFRM and CFRM varying the configuration settings and comparing classification accuracy. We also evaluate the influence of the feature selection algorithms, the inference systems themselves and the quantity of fuzzy sets in the input system.

For each dataset, we performed the preprocessing, transformation and feature creation stages as we described. But, starting at feature selection we performed each stage only on training folds. For example, the fold 1 is used as test fold in classification and evaluation stages, but the remaining folds are used to feature selection and build the combined fuzzy rule base for that fold. The same process is repeated for the rest of the folds and our results are reported as the average of the test folds. Consequently, all kinds of n-grams combined with all transformation techniques described in this work pass to feature selection stage to find out which features would be more fitted to represent documents.

Feature selection algorithms evaluation

To evaluate the feature selection algorithms we start with the following settings: 3 fuzzy sets in the input and CFRM for both datasets. Hence, using the two other parameters unchanged, we can evaluate the feature selection algorithms performance. Besides recall, precision, accuracy and F1 we also verified the average quantity of selected features for each algorithm. Table (1) shows the results from movies and table (2) from Amazon dataset.

Movies	CFS	c4.5
Precision	55.69% ± 8.52%	82.85% ± 20.00%
Recall	79.40% ± 31.15%	37.7% ± 39.16%
Accuracy	53.5% ± 2.16%	55.7% ± 2.46%
F1	59.08% ± 15.84%	35.40% ± 23.04%
Features selected	3.5 ± 0.5	1

Table 1: Results from movie reviews dataset

As we can see, feature selection with c4.5 using CFRM and 3 fuzzy sets in the input obtained better overall precision and

Movies	CFS	c4.5
Precision	70.52% ± 8.91%	63.2% ± 18.45%
Recall	69.8% ± 12.86%	73.8% ± 39.17%
Accuracy	68.75% ± 5.91%	53.5% ± 2.09%
F1	68.79% ± 5.51%	54.56% ± 19.95%
Features selected	6.2 ± 1.66	1

Table 2: Results from Amazon reviews dataset

accuracy in movie reviews dataset, using almost four times less features. However, the inverse occurs in Amazon dataset, where CFS with CFRM performs better than c4.5. But, in Amazon dataset, CFS uses even more features, creating rules with six antecedents, on average, turning the rules less human readable. So, since c4.5 just needed one feature, generating more readable rules, and also considering accuracy as the main reference of performance, despite of less balanced performance showed with lower F1 measure, we decided to use c4.5 in both datasets.

Inference system evaluation

In this subsection we evaluate the performance of the chosen inference systems, CFRM and GFRM. As we did in the last subsection, we fixed the remaining parameters, to better evaluate the inference systems performances, maintaining the c4.5 algorithm and 3 fuzzy sets in the input. Table (3) shows results from movie reviews and table (4) from Amazon dataset.

	CFRM	GFRM
Precision	82.85% ± 20.0%	79.32% ± 15.54%
Recall	37.7% ± 39.16%	45.7% ± 31.71%
Accuracy	55.7% ± 2.46%	60.9% ± 2.55%
F1	35.40% ± 23.04%	48.27% ± 16.01%

Table 3: Inference systems results from movies reviews dataset

	CFRM	GFRM
Precision	63.22% ± 18.45%	65.14% ± 15.51%
Recall	73.8% ± 39.17%	75.7% ± 33.27%
Accuracy	53.5% ± 2.09%	59.65% ± 1.98%
F1	54.56% ± 19.95%	60.97% ± 14.74%

Table 4: Inference systems results from Amazon reviews dataset

The results shows that General Fuzzy Reasoning Method improves accuracy over the Classical Fuzzy Reasoning Method, maintaining feature selection and fuzzy sets unchanged. Also the F1 score shows better balance between precision and recall with GFRM. In this classification task with two classes only, to consider the entire set of rules of a class is a better approach than use only one rule with the highest degree. Hence, GFRM is our choice to achieve better results in this work.

Evaluation of fuzzy sets quantity

Through the last subsections, we have seen the results using 3 fuzzy sets to model our linguistic variables. Following the

decision to pick up c4.5 to reduce the complexity of the rules and make them more human readable, we tried to reduce the fuzzy sets, using only the "Low" and "High" fuzzy sets. Table (5) shows the obtained results to movies and table (6) from Amazon dataset.

	3 fuzzy sets	2 fuzzy sets
Precision	79.32% ± 15.54%	72.09% ± 4.28%
Recall	45.7% ± 31.71%	69.50% ± 8.46%
Accuracy	60.9% ± 2.55%	71.25% ± 4.43%
F1	48.27% ± 16.01%	70.53% ± 5.55%

Table 5: Inference systems results from movie reviews dataset

	3 fuzzy sets	2 fuzzy sets
Precision	65.14% ± 15.51%	73.32% ± 3.08%
Recall	75.7% ± 33.27%	62.5% ± 4.58%
Accuracy	59.65% ± 1.98%	69.9% ± 3.02%
F1	60.97% ± 14.74%	67.43% ± 3.68%

Table 6: Inference systems results from Amazon reviews dataset

Table (5) shows that the accuracy and F1 were significantly improved by removing a fuzzy set, more specifically the "medium", remaining the "low" and "high" fuzzy sets. Also, between movies and amazon datasets, even though very slightly, the best overall results is in movies. This is specially interesting because movie reviews are often reported as the most difficult type of reviews to be classified [Turney, 2002; Pang and Lee, 2004; Chaovalit and Zhou, 2005; Ohana and Tierney, 2009].

In both datasets, the single feature selected by c4.5 was the difference between the sum of positive and negative unigrams and bigrams composed by adjectives and adverbs. With this only feature, we could classify close to 70% of the movies reviews and Amazon reviews with two simple and human readable rules generated by Wang-Mendel method:

- IF the difference between the sum of positive and negative unigrams and bigrams composed by adjectives and adverbs is HIGH then POLARITY is POSITIVE
- IF the difference between the sum of positive and negative unigrams and bigrams composed by adjectives and adverbs is LOW then POLARITY is NEGATIVE

More results

Although we have used the Amazon dataset presented in [Wang *et al.*, 2011] to test and evaluate our work, the evaluation in that paper was related to rating prediction and not classification, making any comparison improper. The same can't be said about the Cornell Movie Reviews. This is a dataset already pre-processed by the authors used in that way for many others papers. Hence, we compare our results with those papers that have used the Cornell Movie Reviews dataset.

Our work can be comparable to [Ohana and Tierney, 2009] and [Taboada *et al.*, 2008] that used strictly the same dataset and they aren't domain dependent as well. They showed

69,35% and 76% of accuracy, respectively. It is important to say that these works do not apply a fuzzy approach and [Taboada *et al.*, 2008] uses many different steps from our work, such as opinion lexicon (they manually created their own), entire different intensifiers set, among others. [Ohana and Tierney, 2009], in the other side, uses many things related to this work, such as Sentiwordnet and many similar and equal documents features.

We can cite others papers that have used a previous version of this movie dataset (that differs in quantity) such as [Ohana *et al.*, 2011] that presented 69,9% of accuracy. Concerning papers that have presented a fuzzy approach, we couldn't find any of them that presented results or anything closely related to this work.

5 Conclusion and further works

This work proposed and evaluated an automated fuzzy opinion mining system to classify the overall sentiment orientation of document's text. Our proposal uses the Wang-Mendel method [Wang and Mendel, 1992] to generate fuzzy rules based on most fitted features among almost 70 features that were extracted and selected from documents. We achieved promising results, reaching 71,25% of accuracy in a 10-fold cross-validation.

Our work is probably the first one to apply Fuzzy Logic and Wang-Mendel method in opinion mining, evidencing results on datasets from previous works. Besides, our results are comparable to previous works that applies non fuzzy techniques. Also, we classified documents with human readable rules using simple fuzzy sets, such as low, high, positive and negative. We contribute as well in the investigation of features that can be relevant to describe and discriminate documents.

We have reported initial results from a ongoing research. As future works, we have many improvements points such as:

- Build a better set of intensifiers and evaluate their influence in final results;
- Improve negation detection and how to better apply it;
- Improve how the fuzzy sets are modeled to inputs from document features;
- Investigate more features that could represent and better distinguish documents;
- Experiment with other feature selection techniques, to investigate the influence of the selected features on fuzzy rules generation.

References

- [Alistair and Diana, 2005] Kennedy Alistair and Inkpen Diana. Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*, 2005.
- [Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
- [Ballhysa and Asilkan, 2012] Elton Ballhysa and Ozcan Asilkan. A fuzzy approach for blog opinion mining-an application to albanian language. *AWERProcedia Information Technology and Computer Science*, 1, 2012.
- [Boucher and Osgood, 1969] Jerry Boucher and Charles E Osgood. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, 8(1):1–8, 1969.
- [Brill, 1995] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.
- [Chaovalit and Zhou, 2005] Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE, 2005.
- [Cintra *et al.*, 2008] Marcos Evandro Cintra, CH de Arruda, and Maria Carolina Monard. Fuzzy feature subset selection using the wang & mendel method. In *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*, pages 590–595. IEEE, 2008.
- [Cordon *et al.*, 1999] Oscar Cordon, Maria Jose del Jesus, and Francisco Herrera. A proposal on reasoning methods in fuzzy rule-based classification systems. *International Journal of Approximate Reasoning*, 20(1):21–45, 1999.
- [Esuli and Sebastiani, 2006] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.
- [Guerini *et al.*, 2013] Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*, 2013.
- [Hall, 1999] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.
- [Karamibekr and Ghorbani, 2012] Mostafa Karamibekr and Ali A Ghorbani. Verb oriented sentiment classification. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, pages 327–331. IEEE, 2012.
- [Kazmier, 2004] Leonard J Kazmier. *Schaum's outline of business statistics*. McGraw-Hill, 2004.
- [Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.
- [Moraes *et al.*, 2012] Rodrigo Moraes, João Francisco Valiati, and Wilson P Gavião Neto. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 2012.
- [Nadali *et al.*, 2010] S Nadali, MAA Murad, and RA Kadir. Sentiment classification of customer reviews based on

- fuzzy logic. In *Information Technology (ITSim), 2010 International Symposium in*, volume 2, pages 1037–1044. IEEE, 2010.
- [Ohana and Tierney, 2009] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.
- [Ohana et al., 2011] Bruno Ohana, Brendan Tierney, and S Delany. Domain independent sentiment classification with many lexicons. In *Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on*, pages 632–637. IEEE, 2011.
- [Pang and Lee, 2004] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.
- [Pang et al., 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- [Quinlan, 1993] RC Quinlan. 4.5: Programs for machine learning morgan kaufmann publishers inc. *San Francisco, USA*, 1993.
- [Quirk et al., 1985] Randolph Quirk, David Crystal, and Pearson Education. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press, 1985.
- [Taboada et al., 2008] Maite Taboada, Kimberly Voll, and Julian Brooke. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*, 2008.
- [Taboada et al., 2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
- [Turney, 2002] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.
- [Wang and Mendel, 1992] L-X Wang and Jerry M Mendel. Generating fuzzy rules by learning from examples. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(6):1414–1427, 1992.
- [Wang et al., 2011] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM, 2011.
- [Wang, 2003] L-X Wang. The wm method completed: a flexible fuzzy system approach to data mining. *Fuzzy Systems, IEEE Transactions on*, 11(6):768–782, 2003.
- [Wiebe, 1990] Janyce M Wiebe. Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 401–406. Association for Computational Linguistics, 1990.
- [Wiebe, 1994] Janyce M Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.
- [Wilson et al., 2004] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769, 2004.
- [Wilson et al., 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.
- [Zadeh, 1965] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.
- [Zadeh, 1996] Lotfi A Zadeh. Fuzzy logic= computing with words. *Fuzzy Systems, IEEE Transactions on*, 4(2):103–111, 1996.