



## An Enhanced MNB Based Model for Explicit and Hidden Sentiment Classification in Imbalanced Datasets

Hajar El Hanhach<sup>1\*</sup>      Mohammed Benkhalifa<sup>1</sup>

<sup>1</sup>*Algorithms, Networks, Intelligent Systems and Software Engineering Research Team,  
 Computer Science Department, Faculty of Science, Mohammed V University Rabat, Morocco*

\* Corresponding author's Email: [elhannach.hajar@gmail.com](mailto:elhannach.hajar@gmail.com)

---

**Abstract:** Sentiment polarity classification (either explicit or hidden) is the process by which information can be extracted to be analysed as positive or negative opinion. Much work on supervised machine learning based sentiment classification has been done considering balanced datasets. However, due to the imbalanced nature of data distribution, sentiment classification becomes a complex task that requires investigating more efficient approach, especially for hidden sentiment. Multinomial Naïve Bayes (MNB) algorithm is one of the most widely used methods in this field, due to its computational efficiency and relatively good predictive performance. However, this classifier performs poorly on imbalanced datasets. In this paper, we propose transformation in MNB prior and conditional probabilities designed to handle explicit and hidden sentiment classification for imbalanced data. To support MNB, we introduce an original frequency model based on using synonym and antonym semantic relations. Our approach is empirically evaluated according to: (1) its impact on MNB training data quality for minority classes, (2) its comparison to similar works involving MNB for imbalanced datasets and (3) its comparison to commonly used classifiers for sentiment classification. The experimental results show that our model improves training data quality and therefore helps MNB boost its performance to achieve the best results for explicit and hidden imbalanced datasets.

**Keywords:** Hidden sentiment, Machine learning, Imbalanced datasets, Algorithmic level.

---

### 1. Introduction

Several existing works on supervised learning based sentiment analysis assume the balance between negative and positive data [1, 2]. However, in several cases, sentiment polarity classification suffers of imbalanced class distribution wherein samples of one class outnumber the other one. The class with more samples is known as majority class and the class with fewer data is the minority [3]. For the imbalanced sentiment classification, the complication arises when the number of data representing the class of interest are the lower. Overall, for product and restaurant datasets, the positive opinions are widely the majority class while the negative ones, which are mainly used in product improvement process, are the minority class. Therefore, without class imbalanced solutions, the

supervised classification is overwhelmed by the majority class and the results can be of no use.

The imbalanced classification solutions fall into three main categories: data, data-algorithmic and algorithmic levels. The data level is the pre-processing techniques that aim to attain better input data. The re-sampling methods are the most popular techniques in this category. In particular, the Synthetic Minority Oversampling Technique (SMOTE) [4] which is commonly combined with machine learning based classification to balance class distribution [5, 6]. For the data-algorithmic level, we found the Cost sensitive learning approach. The cost-based model represents the misclassification cost of classifying a sample from class  $i$  as class  $j$ . The cost sensitive learning had proven its efficiency for imbalanced datasets problem, yet it barely used. Indeed, the cost matrix

building is a challenging task that requires expert's opinions or learning from others approaches [7]. The third way that deals with imbalanced datasets is the algorithmic level. This level comprises ensemble methods [8] and the algorithmic classifier modification [9]. The ensemble methods design the combination of multiple classifiers to improve final obtained performances. The algorithmic classifier modification attempts to enhance the learning ability of supervised classifier to boost classification performances for imbalanced datasets.

This study, which is algorithmic based, is directed toward improving MNB classifier to allow better classification performances for explicit and hidden sentiment classification with imbalanced data distribution. The proposed model operates on two major steps to enhance training data using semantic relations from WordNet (WN) along with a new frequency model. At the pre-processing, WN and dependency parser are based to expand vocabulary using synonym and antonym relations. Two semantic relations are exploited differently in this approach. At the first step, synonyms are considered to enhance training data for majority and minority classes. For hidden sentiment extraction, we investigated the findings shown in [10] that consider different synonym and definition subsets for adjective, adverb and verb. At the second step, the antonyms are used to support the minority class based on their distribution on the majority class. The enhanced model supports the class prior probability and probability distribution over vocabulary of MNB to avoid the bias of the training toward the majority class and support minority class identification.

Most of the effort in this study aim at improving the F1-score for hidden and explicit sentiment polarity classification which is severely affected by the imbalanced distribution. The developed model is applied on different benchmark datasets and compared to SVM and KNN classifiers, and two approaches which involve SMOTE technique differently combined with MNB. The first one [11] is based on pre-processing techniques to enhance the MNB classifier performances, and the second one [12] addresses the sentiment analysis using MNB coupled with semantic relations from WN, SentiWordnet (SentiWN).

The overall structure of this study takes the form of four main sections organised as follows. In section 2, related works are reviewed followed by preliminaries in section 3. Section 4 is dedicated to the problem statement, and in section 5, the proposed approach is presented in details. The experimental setting adopted is exposed and the

obtained results are presented and discussed in section 6. Finally, conclusions and future work are presented.

## 2. Related works

The survey conducted by Haixiang et al [13] shows that there is a considerable number of published studies developed to cope with imbalanced data classification. Chawla et al. in [4] introduce the SMOTE technique. This method is an oversampling technique based upon the generation of synthetic samples for minority class rather than a simple oversampling with replacement. The synthetic instances generation depends on two main factors (1) The amount of oversampling required, and (2) The K-minority class nearest neighbours.

A number of works have examined the impact of k-Nearest Neighbour and provide a fuzzy approach [14, 15] while others [11, 16] had studied the impact of SMOTE on different classifier performances. Combining MNB in one hand, and SVM, in the other hand, with SMOTE is a widely applied method to sentiment classification for imbalanced datasets. In [11], authors have focused their research on sentiment classification applied on imbalanced data extracted from Twitter and Facebook. The classification is performed using SMOTE technique combined first with MNB and then with SVM. The experiments have been conducted in order to compare performances of the two classifiers, and to determine factors involving misclassification problems of using SMOTE, particularly, in pre-processing and test phases. The obtained results have shown that class distribution is among factors that negatively affect SMOTE technique. To alleviate the impact of class distribution and evade the overfitting problems, two steps are performed. The randomization technique is applied at the pre-processing phase, and the 10 cross validation is considered rather than the percentage splitting of '70:30' for the test. Their final obtained results showcase the importance of their method to enhance classification performance using SMOTE while dataset does not contain a large scale of noisy terms. However, noise is an unavoidable issue that affects data collection and preparation. In attempt to defuse the impact of noisy terms, several works had investigated the WN and SentiWN as lexical components used at the feature selection process [12, 17]. SentiWN attributes scores representing the sentiment polarity orientation to each word. The given score aims to determine the relevant terms with sentiment. The work of [12] considered SentiWN for sentiment analysis. The authors

developed two systems to deal with two sentiment analysis tasks applied to twitter messages. The first task, known as Task A, classifies the sentiment polarity of one target word, while the second task, Task B, classifies the polarity of the whole tweet. Task A relies on Bag of words and feature expansion methods. The authors expand upon the training data through WN Synonym and Definition relations coupled with SentiWN to generate a new weighted vocabulary. Definition relation is used to predict the different meaning of the same word and then choose the appropriate SentiWN score for each definition. Task B computes the number of positive and negative terms appearing in message. After applying the two main pre-processing steps, which are stop-words removal and stemming, they remove terms that occurred only once. The next step of this task computes the number of positive and negative terms using three lexicon resources: SentiWN, Polarity Lexicon and the General Inquirer. For the two tasks, the classification is done using MNB in comparison with SVM. For Task A, MNB achieves the best results while SVM performed better for Task B. Our proposed method is closed to [12] as it used WN semantic relations, namely synonym and definition for feature expansion, but we introduce a new frequency model for weighting vocabulary rather than SentiWN. Furthermore, instead of using SMOTE, the proposed model promotes the number of minority documents by considering semantic relations between the positive and negative sentiment classes.

### 3. Preliminaries

Multinomial Naive Bayes (MNB) is the version of Naive Bayes that is commonly used for text categorization and sentiment classification. This is due to its simplicity, computational efficiency and relatively good predictive performance. For a given document  $di$ , the probability that  $di$  belong to class  $cj$  is computed as

$$\Pr(di/cj) = \log(P(cj)) + \sum_{i=1}^{|V|} fi \log(\Pr(Ti/cj)) \quad (1)$$

Where  $P(cj)$  is the prior probability of class  $cj$ ,  $fi$  is the number of occurrences of a term  $Ti$  in document  $di$ , and  $\Pr(Ti/cj)$  is the conditional probability of term  $Ti$  in class  $cj$ .

#### A. Prior probability

The prior probability distribution of class  $cj$ , denoted by  $P(cj)$ , is a weight that indicates the

relative frequency of class  $cj$ . The  $P(cj)$  is obtained as follows:

$$P(cj) = \frac{Ncj}{Nc} \quad (2)$$

Where  $Ncj$  stands for the number of documents in class  $cj$ , and  $Nc$  denotes the number of all documents.

#### B. Conditional probability:

The MNB classifier is based on the conditional probability to represent the probability distribution over vocabulary  $\Pr(Ti/cj)$ . The conditional probability  $\Pr(Ti/cj)$  of term  $Ti$  is based on the number of occurrences of this word in training data  $count(Ti, cj)$ .

$$\Pr(Ti/cj) = \frac{count(Ti, cj)}{count(cj)} \quad (3)$$

### 4. Problem statement

For MNB, complications arise for sentiment analysis applied on imbalanced datasets. The MNB classification leads to favor the majority class and shows poor classification rates on minority one. The Bayesian analysis implies that a new document is more likely to belong to the class with a higher prior probability than others. In the case of imbalanced datasets, the number of documents in minority class is very small compared to the majority class. This results in very low values of  $P(cj)$  for the minor class. Accordingly, the prior probability has a substantial weight that markedly biases the identification of minority class and leads the classifier to favor the majority one. For the conditional probability, the minority class instances do not provide sufficient information for training. Consequently, considering the number of occurrences only is insufficient in such situations.

Hence, to support MNB sentiment classification for imbalanced datasets we provide a new approach that enhances prior and conditional probabilities of MNB.

### 5. The proposed method

We proposed an enhanced model based on frequency functions and WordNet to alleviate the effect of prior and conditional probabilities.

### 5.1 Prior probability transformation

We improve the prior probability for minority class through the use of WordNet semantic relations, namely Synonym and Antonym, between terms implying negative and positive sentiments.

$$\text{For majority class } P(c_j) = \frac{Nc_j}{Nc} \tag{4}$$

$$\text{For minority class } P(c_j) = \frac{Nc_j + NA_j}{Nc}$$

Where  $NA_j$  stands for the number of documents from the majority class  $c_j$ . The considered documents contain antonyms of sentiment terms representing documents in minority class  $c_j$ .

### 5.2 Conditional probability transformation

In order to support MNB conditional probability, we introduce the frequency function  $F(Ti)$  based on: (1) The WN frequency of the extracted related terms of  $Ti$  denoted by  $f_{wn}(Ti)$ , (2) The training data frequency of  $Ti$   $ft(Ti)$ , and (3) The indicator factor  $\partial(Ti)$ . The new frequency function  $F(Ti)$  is used on the new conditional probability equation as follows:

*For majority class*

$$Pr(Ti/c_j) = \frac{F(Ti, c_j) + \sum_k^{Sn} F(S_k, c_j)}{count(c_j)}$$

*For minority class* (5)

$$Pr(Ti/c_j) = \frac{F(Ti)}{count(c_j) + count(Ai, c_j)}$$

$$\text{With } F(Ti) = F(Ti, c_j) + \sum_k^{Sn} F(S_k, c_j) + \sum_k^{An} F(A_k, c_j)$$

Where  $S_k$  and  $A_k$  stand for the  $k^{th}$  Synonym of term  $Ti$  and the  $k^{th}$  antonym of term  $Ti$  appearing in class  $j$  respectively.  $Sn$  and  $An$  denotes, in that order, the number of synonyms and antonyms of term  $Ti$ .

The new considered Eq. (4) and Eq. (5) are detailed in the following section.

### 5.3 The enhanced MNB Model for sentiment classification

As shown in Fig.1, the proposed model operates on three major phases: Sentiment term extraction, Training data enhancement using Synonyms and the Training data enhancement for minority class using Antonyms. The first phase creates for each extracted sentiment term  $Ti$  two lists: the first list contains  $Ti$

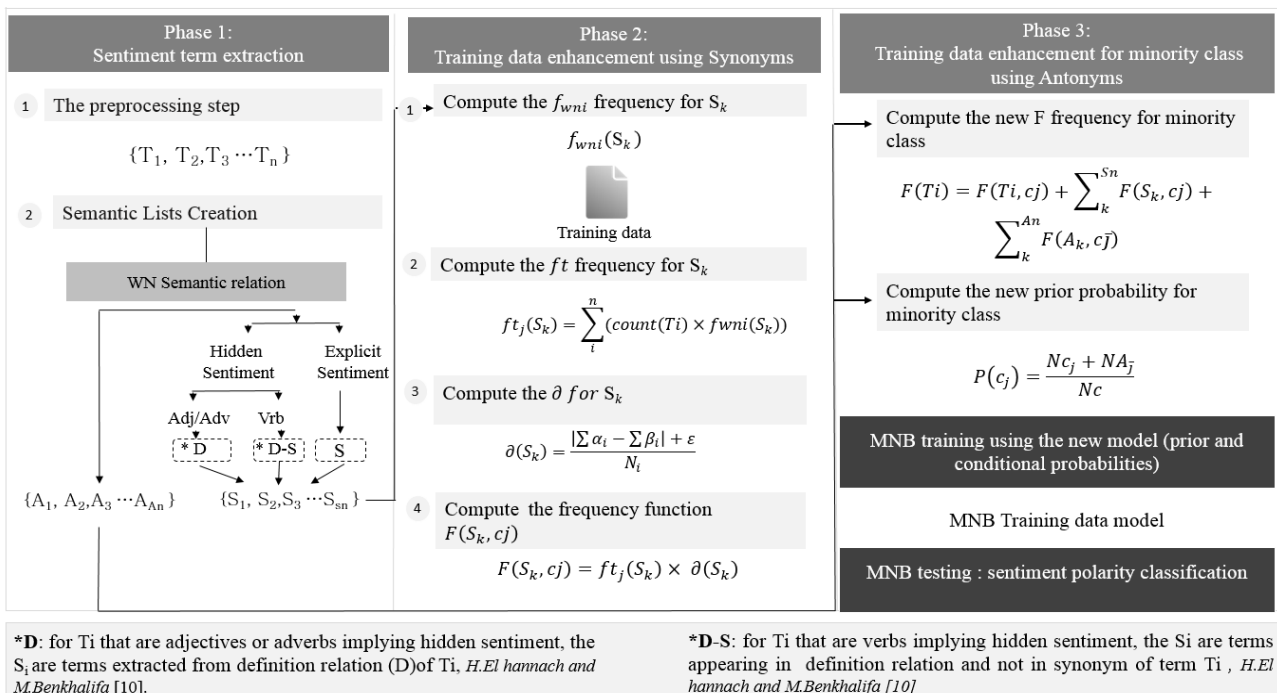


Figure. 1 Summary of our approach

and its synonyms while the second presents its Antonyms. The second phase computes the frequency function  $F(S_k, cj)$ , used in Eq.(5), for each Synonym of Term  $T_i$  in class  $cj$ . The second frequency  $F(A_k, cj)$ , shown in Eq.(5), the prior and conditional probabilities are computed in the third phase. At last, the MNB training model is generated based on the two enhanced Eq. (4) and Eq. (5).

### Phase 1: Sentiment terms extraction

Sentiment terms extraction consists of pre-processing and semantic lists creation. The two tasks rely on POS tagger and WordNet to extract the potential sentiment words  $T_i$  and their WN extracted terms  $S_i$  and  $A_i$  to represent each corpus document.

*The pre-processing:* For each document, a dependency parser is applied to extract a list of adjectives, verbs and adverbs. The extraction process is followed by stemming and noisy terms removal. As result, a list of potential sentiment terms  $T_i$  is obtained  $\{T_1, T_2, \dots, T_n\}$ .

*Semantic Lists Creation:* It concerns the two semantic lists generation. For each  $T_i$ , Synonyms and Antonyms extracted from WN are listed in  $S_i$  and  $A_i$  respectively.

List of  $T_i$  and its synonyms, where  $S_1 = T_i$

$$S_i = \{S_{i_1}, S_{i_2}, \dots, S_{i_{S_n}}\}.$$

List of Antonyms of  $T_i$ ,

$$A_i = \{A_{i_1}, A_{i_2}, \dots, A_{i_{A_n}}\}.$$

For hidden sentiment, the  $S_i$  extraction is based on the findings shown in [10] that consider different synonym and definition subsets for adjective, adverb and verb. For adjective and adverb, the synonyms used are extracted from the definition relation (D) while the part of synonyms appearing in D contains more reliable terms with closer meaning than terms appearing in Synonym relations [10]. For the potential verbs that imply hidden sentiment we extract terms appearing only in definition and not in synonyms (D-S). Verbs have many synonyms with different contexts that lead to poor results when considering all of them [10].

### Phase 2: Training data enhancement using Synonyms

This phase operates on 4 steps. It relies on the extracted lists of synonyms and the frequency model to compute  $F(S_k, cj)$  for each Term  $T_i$ .

In step 1, the WN frequency function  $f_{wni}(S_k)$  is computed for  $S_k$ . This frequency represents the occurrences of  $S_k$  as a synonym of  $T_i$  in WN. It distinguishes between frequent and seldom synonyms of  $T_i$ . The WN frequency is considered as indication reliability of the synonym term  $S_k$  to  $T_i$ .

In step 2, we aim to introduce a context domain information extracted from training data. The first information considered is the class training data frequency  $ft_j$ . This frequency is computed for each  $S_k$  combining its  $f_{wni}$  (as synonym indicator reliability for  $T_i$ ) with  $T_i$  occurrence in class  $j$ .

$$ft_j(S_k) = \sum_i^n (count(T_i) \times f_{wni}(S_k)) \quad (6)$$

Where  $count(T_i)$  represents the number of occurrences of  $T_i$  in class  $j$ .

The frequency  $ft_j$  intends to highlight relevant  $S_k$  terms with high scores. Those  $S_k$  that appear in multiple  $T_i$  representations have more chance to be potential sentiment words representing class  $j$ .

In step 3, the aim is to reduce the computed  $ft_j$  score for irrelevant terms. To distinct between irrelevant noisy and potential sentiment words, the second context domain information is used. For this step, we introduce the indicator factor function  $\partial(S_k)$ . The  $\partial(S_k)$  penalizes words appearing twice at negative and positive classes. The  $\partial(S_k)$  is calculated as follows:

$$\partial(S_k) = \frac{|\sum \alpha_i - \sum \beta_i| + \varepsilon}{N_i} \quad (7)$$

Where  $\alpha_i$  and  $\beta_i$  are respectively the numbers of occurrences of  $S_k$  in minority and majority classes.  $N_i$  is the number of occurrences of  $S_k$  and  $\varepsilon$  is used to avoid the case of  $\partial(S_k) = 0$  when  $\sum \alpha_i = \sum \beta_i$ . The best empirical value of  $\varepsilon$  (which is 0.3) is retained.

Finally, the frequency function  $F(S_k, cj)$  is computed in step 4 as follows:

$$F(S_k, cj) = ft_j(S_k) \times \partial(S_k) \quad (8)$$

The frequency  $F(S_k, cj)$  which is based on synonym relation supports minority and majority class by considering high weights for relevant sentiment

terms. However, despite the use of the frequency function, majority class with its large number of documents dominate the classification. With the few number of documents, prior and conditional probabilities induce misclassification for minority class. In attempt to handle this problem, antonyms are used to promote the number of documents in minority class in order to enhance the prior probability and to support the frequency function.

### Phase 3: Training data enhancement for minority class using Antonyms

In this phase, the proposed approach takes advantage of the Antonym semantic relation between minority (negative opinion) and majority (positive opinion) classes. In sentiment analysis, it's obvious to find antonyms of terms of the minority class used in majority class documents, e.g. if 'bad/hate/awful' are used in negative class, it's more likely to find 'good/like/attractive', in the positive one. Using those antonyms, we support minority class recognition as follows.

For each sentiment term  $Ti$  of the minority class, we extract documents that contain  $Ti$  antonyms from majority class. The extracted documents are considered both as minority and majority documents. The consideration of those documents for the minority class  $cj$ , implies that, the frequency of the term  $Ti$ , as given in Eq.(9), is equal to : the frequency of  $Ti$  plus the sum of the frequency of their synonyms in  $cj$  ( $\sum F(S_k, cj)$ ) plus the sum of frequency of  $Ti$  antonyms in the extracted document denoted by  $\sum F(A_k, c\bar{j})$ .

$$F(Ti) = F(Ti, cj) + \sum_k^{Sn} F(S_k, cj) + \sum_k^{An} F(A_k, c\bar{j}) \quad (9)$$

The new prior probability of class  $cj$  is computed considering the new added document as follows:

$$P(c_j) = \frac{Nc_j + NA_j}{Nc} \quad (10)$$

At last, MNB training is performed using the prior and conditional probabilities according to Eq. (4) and Eq. (5).

## 6. Experiments and results

This section describes the experimental design adopted to evaluate the proposed approach. It

comprises the pre-processing techniques, the datasets chosen, the classifiers utilized, the performances evaluation metrics considered, the experimental protocols and finally the obtained results with the discussion.

### 6.1 Experimental setup

**The pre-processing.** At this step, filtering text techniques along with the potential sentiment terms extraction process are applied to generate a list of relevant terms without noisy content.

**Datasets.** All experiments were performed on six electronic products corpora and one restaurant corpus. Each corpus used is composed of reviews with explicit and hidden sentiments. Electronic products datasets contain five corpora provided by [18], extracted from Amazon and prepared for implicit aspect extraction. The five datasets include more than 3k sentences from which we construct the datasets used in this work. The laptop is the last electronic product dataset which contains 3864 sentences with explicit and hidden sentiments. The last corpus is the restaurant dataset (denoted by Rest.) used in [19] with 3k documents. The number of instances and the imbalanced ratio (IR), shown in Eq. (11), of each dataset are shown in Table 1.

**Classifiers.** In addition to MNB classifier, the proposed approach is compared to Support Vector Machine (SVM) and the K-nearest neighbours (KNN) classifiers.

1) *The Support Vector Machine:* The support vector machine SVM had proven its accomplishment in text classification and sentiment analysis issues. Several studies

Table 1. Imbalanced ratio of explicit and hidden datasets

	Explicit datasets		Hidden datasets	
	Number of Instances	IR	Number of Instances	IR
<b>Apex</b>	374	4.1	202	1.1
<b>Canon</b>	430	13.3	146	2.8
<b>Jukebox</b>	1145	10.1	219	2.1
<b>Nikon</b>	271	1.6	158	5.4
<b>Nokia</b>	584	8.4	163	4.2
<b>Laptop</b>	3087	5.3	349	1.1
<b>Rest.</b>	2824	3.6	220	1.5

$$IR = \frac{Nbr\ of\ instances\ of\ majority\ class}{Nbr\ of\ instances\ of\ minority\ class} \quad (11)$$

compared their obtained results to SVM in order to validate their findings. In this work, the version Sequential Minimal Optimization (SMO) developed in [20] is considered using the polynomial Kernel and the parameter  $C=1$ .

2) *The K-Nearest Neighbors*: KNN is a lazy learning classifier, that stores training examples directly from training instances to use them to determine the class of a new instance. KNN is widely applied in sentiment classification and aspect extraction and had markedly achieved good performances for imbalanced datasets classification [14].

**The evaluation measure.** The machine learning techniques and evaluation measures adopted in this work are performed using Weka platform [21]. All the experiments are carried out using the 10-fold cross-validation to decrease the uncertainty of data split between training and test data. To assess the effect of the proposed approach, we use the F1-score which is commonly adopted for evaluation in such datasets. The F1 represents the equally weighted average of recall and precision measures as stated in Eq. (12).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100 \quad (12)$$

## 6.2 Experimental protocols

The evaluation is performed according to the three following points:

### 1) *Training data enhancement for minority class using Antonyms*

To assess the impact of considering Antonyms in training data enhancement for minority class, we conduct the two following experiments:

- Our Approach without antonyms: it refers to the use of MNB with WN Synonym relation and frequency function  $F(S_k, cj)$  defined in phase 2 of our approach.
- Our Approach: it denotes the experiment performed using the whole proposed method.

### 2) *Comparing our approach to similar works:*

The second point pertains to comparing our approach to MNB using the SMOTE technique. The SMOTE promotes the number of documents of the minority class by creating synthetic samples based upon the existing minority documents. The SMOTE

process consists of identifying the feature vector of each instance and its nearest neighbours and computing the linear distance between instance and each neighbour. Afterwards, the difference is multiplied with a random number between 0 and 1. Finally, the obtained random value is added to the feature vector in order to create the new synthetic sample along the line segment between the instance and its neighbour. The final obtained synthetic samples augment the number of documents of the minority class, and therefore support the prior and conditional probabilities of MNB classifier. We compare our method to each one of the following experiments using MNB with SMOTE technique:

- $MNB_{(Randomization + SMOTE)}$ : we applied the method of [11], which is based on randomization technique and SMOTE, to products and restaurants datasets used in this work.
- $MNB_{(WN + SentiWN + SMOTE)}$ : we consider the proposed method presented in [12] based on WN and SentiWN for feature expansion. We combine the proposed method with SMOTE for sentiment classification of products and restaurants datasets used in this work.

### 3) *Comparing our approach to other classifiers:*

So far, we had only considered MNB coupled with SMOTE for comparison. In order to compare the effectiveness of our approach to other classifiers using SMOTE and evaluate the impact of our frequency function on training data enhancement, the F1-performances of SVM, KNN and our approach are compared. SVM and KNN are commonly used techniques for classification and more particularly in sentiment analysis. SVM and KNN are applied to sentiment polarity classification using two pre-processing techniques for training data enhancement that allow us to examine : (1) if our frequency function is also supportive for other classifiers than MNB and (2) if our frequency model allows classifiers achieve better performances than SentiWN. For this evaluation, two sets of experiments are performed.

At the first set of experiments, we applied a widely used method in sentiment analysis [12]. We extract synonyms from WN, then the opinion polarity information of each extracted term is given by SentiWN. The final obtained model is based on the training process with SMOTE technique. For conciseness, the two experiments are denoted by  $SVM_{+(1)}$  and  $KNN_{+(1)}$ .

At the second set of experiments, we extract synonyms from WN, then the opinion polarity

Table 3. F1-score of our approach with and without Antonyms for explicit and hidden sentiment classification

<b>EXPLICIT DATASETS</b>	Apex	Canon	jukebox	Nikon	Nokia	Laptop	Restaurant
(1) Our Approach without antonym	88.4	87.2	87.3	79.1	90.9	80.7	86.7
(2) Our Approach	<b>94.2</b>	<b>90.6</b>	<b>90.6</b>	<b>91.1</b>	<b>97.6</b>	<b>97.7</b>	<b>92.2</b>
Average Improvement Rate (1) / (2)	6.6%	3.9%	3.7%	15.1%	7.4%	21.1%	6.3%
<b>HIDDEN DATASETS</b>	Apex	Canon	jukebox	Nikon	Nokia	Laptop	Restaurant
(1) Our Approach without antonym	85.6	85.7	88.44	86.1	87.7	91.7	92.8
(2) Our Approach	<b>85.8</b>	<b>85.7</b>	<b>90.1</b>	<b>86.8</b>	<b>88.4</b>	<b>92.5</b>	<b>93</b>
Average Improvement Rate (1) / (2)	0.2%	0%	2%	0.9%	0.7%	0.9%	0.2%

Table 2. Comparing F1-score of Our approach with other methods for explicit and hidden sentiment classification

<b>EXPLICIT DATASETS</b>	Apex	Canon	jukebox	Nikon	Nokia	Laptop	Rest
[11] MNB+Randomization +SMOTE	70.8	60.7	77.9	70.1	74.5	63.9	78.5
[12] MNB+WN+SentiWN+SMOTE	85	72.4	84.4	79	81.7	74.9	83.4
(3) Our Approach	<b>94.2</b>	<b>90.6</b>	<b>90.6</b>	<b>91.1</b>	<b>97.6</b>	<b>97.7</b>	<b>92.2</b>
Average Improvement Rate (11) / (3)	33%	49.2%	16.3%	29.9%	31%	52.8%	17.4%
Average Improvement Rate (12) / (3)	10.9%	25.2%	7.4%	15.3%	19.5%	30.4%	10.6%
<b>HIDDEN DATASETS</b>	Apex	Canon	jukebox	Nikon	Nokia	Laptop	Rest
[11] MNB+Randomization +SMOTE	59.1	65	69	68.5	62.3	83.4	68.7
[12] MNB+WN+SentiWN+SMOTE	63.4	77.3	74.4	71.5	73.2	82.6	83.2
(3) Our Approach	<b>85.8</b>	<b>85.7</b>	<b>90.1</b>	<b>86.8</b>	<b>88.4</b>	<b>92.5</b>	<b>93</b>
Average Improvement Rate (11) / (3)	45.1%	31.8%	30.5%	26.7%	41.8%	10.9%	35.3%
Average Improvement Rate (12) / (2)	35.3%	10.9%	9.3%	21.5%	20.8%	12%	11.8%

information of each extracted term is given using our proposed Frequency Function  $F(S_k, c_j)$ . The final obtained model is based on the training process with SMOTE technique. The two experiments are denoted by SVM<sub>+(2)</sub> and KNN<sub>+(2)</sub>.

### 6.3 Results and discussion

The experiments results are presented in this section according to the three points mentioned above.

#### 1) Training data enhancement for minority class using Antonyms

Table 2 provides the F1-scores with the average improvement rates of our approach with and without considering training data enhancement using antonyms (phase 3 of our approach) applied in explicit and hidden corpora. The use of our approach reaches good F1-scores for both datasets. For explicit and hidden data, the average F1-score is about 93.4% and 88.8% respectively. To evaluate the impact of considering antonyms for our approach, we compare (1) to (2). Table 2 shows that there is a clear trend of increasing performances for

sentiment classification. The use of antonym frequency function  $F(A_k, c_j)$  for the conditional probability and considering antonym documents in prior probability support MNB to detect the minority class and therefore better represent it. The average improvement rates of ((1)/(2)) proved that the integration of antonyms seems to be more beneficial for explicit sentiment classification (+9.2%) than for hidden one (+0.7%). Overall, the difference between the two results can be explained by the use of terms with opposite semantic orientations to express explicit opinion. For example, to love/to hate, inexpensive/ expensive... However, antonyms are uncommonly used to imply a hidden sentiment. In the sentence 'it breaks easily' the verb to break implies a negative hidden opinion although 'to repair' does not suggest any positive sentiment. Hence, Training data enhancement for minority class using Antonyms allows greatest performances when the vocabulary used involves words with opposite meanings.

#### 2) Comparing our approach to similar works



From the obtained results shown in Table3, we observe that our approach achieves the highest performances for all explicit and hidden datasets. Those obtained results prove that the transformation proposed in this work helps MNB to separate both classes and highlight relevant terms that imply or explicitly express a sentiment.

Comparing to [11], the average improvement rates of our approach is about +32.8% and +31.7% for explicit and hidden datasets respectively. Although, the integration of SMOTE technique supports MNB prior and conditional probabilities, it does not help MNB to outperform our approach. There are two possible reasons to explain this result.

The first one can be related to the lack of lexicon. The [11] is based on randomization and SMOTE to support MNB for sentiment classification without considering any lexicon resources which can hardly be helpful for sentiment classification in such data, in particular, for hidden datasets where [11] achieves its worst performances. However, our approach relies on WN semantic relations to expand the minority training data. The second reason can be related to noisy data. The performances of SMOTE oversampling are powerfully influenced by the class distribution and the quality of data. Due to the lack of information of minority class, it's difficult to distinguish between relevant and noisy terms in [11]. However, in our approach, the scores attributed to potential extracted sentiment words and their WN extracted terms are computed considering their distribution on minority and majority classes. The use of semantic relation between terms of positive and negative classes, namely Antonym relation, compensated the lack of information by supporting terms by their opposites that occur in majority class. Hence, the final computed scores indicate reliability for a given term in sentiment representation according to specific context.

Comparing to [12], the average performance improvement rates of our approach are about 17% and 19 % for explicit and hidden datasets respectively. The integration of lexicon resources with the SMOTE in [12] significantly increases performances for all datasets in comparison to [11], even so there are some limitations in [12]. The integration of SentiWN, which does not consider contextual polarity of words, assigns the same score to words across diverse contexts. In contrast, our approach is domain specific polarity oriented technique. Through the use of the frequency function  $F(T_i)$ , our approach constructed its own model to weight sentiment terms with a positive or negative score according to the corpus context. Therefore, the built model is more supportive for

sentiment classification. Mainly, for hidden datasets where SentiWN achieves its worst performances. Besides,  $F(T_i)$  function allows the distinction between two sentiment terms appearing in the same test document. The term with higher  $F(T_i)$  score is more appropriate to be the term with the sentiment. This distinction is helpful for documents with negation. For those documents, the negation is considered if it is applied to term with high  $F(T_i)$  score. For example, in the sentence '*because of this, I would not recommend this small toy*' the negation form is considered while the verb '*to recommend*' had a higher  $F(T_i)$  score than the adjective '*small*'. However, in '*if t-mobile does not offer such this powerful phone most likely I will get the t610*' the negation is ignored due to the high score of the adjective '*powerful*' in comparison to the verb '*to offer*'.

### 3) Comparing our approach to other classifiers

Figs. 1 and 2 show F1 performances of SVM, KNN and our approach on explicit and hidden sentiment classification respectively. SVM+(1) and KNN+(1) stand for SVM and KNN using WN and SentiWN coupled with SMOTE technique. While SVM+(2) and KNN+(2) denote experiments considering WN with the proposed frequency function and SMOTE.

The first set of analysis, of Fig 1 and 2, examines the impact of our frequency function  $F(S_k, c_j)$  on SVM and KNN performances in comparison with SentiWN. For all datasets, the two classifiers achieve their best scores when considering the frequency function as the weighting model. For explicit datasets, the average improvement rates for SVM and KNN are respectively about +9.2% and +7.5% for explicit data and +20.4% and +21.9% for hidden corpora. Those results comply with our findings at the second point and prove that  $F(S_k, c_j)$  is also helpful for other classifiers than MNB. KNN classifier is known to be sensitive to bad features that augment the class representation dimension. Using  $F(S_k, c_j)$ , KNN is able to reduce the dimension by removing irrelevant terms with lower  $F(S_k, c_j)$  scores. For the SVM,  $F(S_k, c_j)$  eliminates the effect of terms appearing both in positive and negative classes. This prevents to blur the borderline between classes.

SVM is often considered as the classifier that makes the greatest outcomes for sentiment classification. For explicit datasets, SVM outperforms our approach in Jukebox (+6.9%) and Nikon (+5.3%) while the proposed model provides

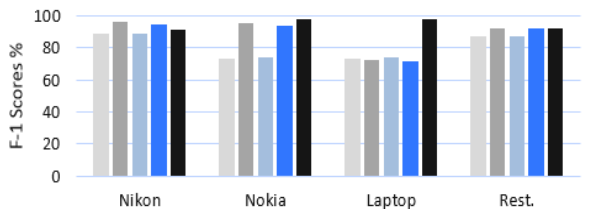
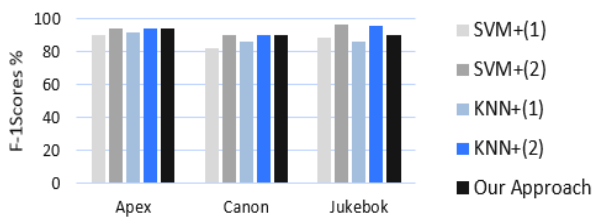


Figure. 2 F1-score of SVM, KNN and our approach in explicit datasets

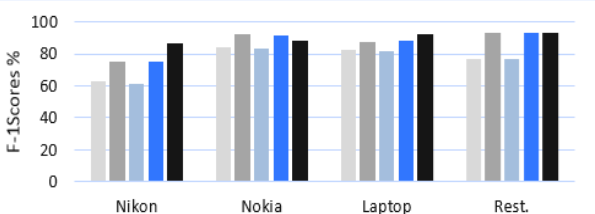
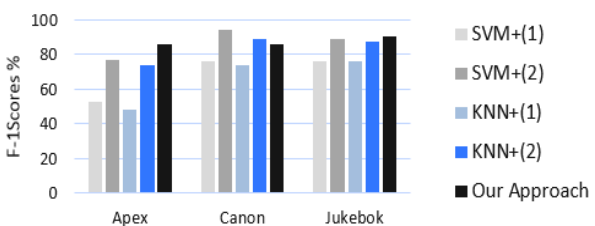


Figure. 3 F1-score of SVM, KNN and our approach in hidden datasets

better performances in the five others datasets (+9.3%). For hidden datasets, our approach supplies worst results than of SVM (-9.2% for Canon and -4.2% for Nokia). For the other five datasets, our method provides better performances (The average improvement rate is about +8.6%). Overall, these results attest that our enhanced MNB model achieves comparable performances to those of SVM and KNN in sentiment classification for imbalanced datasets.

### 7. Conclusion

This paper proposes a transformation of MNB prior and conditional probabilities to deal with explicit and hidden sentiment classification for imbalanced datasets. The proposed rely on WN semantic relation and context domain information to generate a new frequency function supporting MNB probabilities.

The F1-scores acquired in this work for product and restaurant datasets prove the effectiveness of

this approach. The method performs significantly better when considering antonym sematic relation for training data enhancement for the minority class, namely for explicit datasets where the improvement is about +9.2%. Comparing to similar works supporting MNB by SMOTE technique, our approach had proven its ability to improve MNB performances through the use of WN and frequency function to enhance training data. Furthermore, the evaluation results comparing our approach to others classifiers using SMOTE demonstrated that the proposed model yields good results as those obtained by SVM and KNN supported by our frequency function and achieves better performances than those obtained by the two classifiers using SentiWordNet.

Further work will investigate the problem of multiclass classification with imbalanced data. Through the use of a multiple WN semantic relations the proposed method will be extended to deal with the multiple granularity aspect based sentiment analysis with imbalanced data.

### 8. References

- [1] W. Maharani, "Microblogging sentiment analysis with lexical based and machine learning approaches", In: *Proc. of International Conf. of Information and Communication Technology*, pp. 439–443, 2013.
- [2] D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification", In: *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*, pp. 1422–1432, 2015.
- [3] D. D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: A Review", *International Journal of Computing and Business Research*, Vol. 5, No. 4, p. 1-15, 2014.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.
- [5] B. Krawczyk, B. T. McInnes, and A. Cano, "Sentiment Classification from Multi-class Imbalanced Twitter Data Using Binarization", In: *Proc. of International Conf. Hybrid Artificial Intelligence Systems*, pp. 26–37, 2017.
- [6] F. Wu, C. Wu, and J. Liu, "Imbalanced Sentiment Classification with Multi-Task Learning", In: *Proc. of the 27th ACM International Conf. on Information and*

- Knowledge Management - CIKM '18*, pp. 1631–1634, 2018.
- [7] B. Krawczyk, M. Woźniak, and G. Schaefer, “Cost-sensitive decision tree ensembles for effective imbalanced classification”, *Applied Soft Computing*, Vol. 14, pp. 554–562, 2014.
- [8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 42, No. 4, pp. 463–484, 2012.
- [9] B. Lerner, J. Yeshaya, and L. Koushnir, “On the Classification of a Small Imbalanced Cytogenetic Image Database,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 4, No. 2, pp. 204–215, 2007.
- [10] H. El Hannach and M. Benkhalifa, “WordNet based Implicit Aspect Sentiment Analysis for Crime Identification from Twitter”, *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 9, No.12, 2018.
- [11] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, “An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set”, In: *Proc. of International Conf. on Engineering, Applied Sciences, and Technology (ICEAST)*, pp. 1–4, 2018.
- [12] H. Poursepanj, J. Weissbock, and D. Inkpen, “uOttawa: System description for SemEval 2013 Task 2 Sentiment Analysis in Twitter”, In: *Proc. of the Second Joint Conf. on Lexical and Computational Semantics (\*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation*, pp. 380–383, 2013
- [13] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications”, *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.
- [14] H. Patel and G. S. Thakur “Classification of Imbalanced Data Using a Modified Fuzzy-Neighbor Weighted Approach”, *International Journal of Intelligent Engineering and Systems*, Vol. 10, No. 1, pp. 56–64, 2017.
- [15] N. G. Siddappa and T. Kampalappa “Adaptive Condensed Nearest Neighbor for Imbalance Data Classification”, *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 2, pp. 104–113, 2019.
- [16] P. Sarakit, T. Theeramunkong, and C. Haruechaiyasak, “Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm”, In: *Proc. of the 2nd International Conf. on Advanced Informatics: Concepts, Theory and Applications*, pp. 1–5, 2015.
- [17] C. Pandi, T. S. Dandibhotla, and V. V. Bulusu, “Corpus Linguistic Rules Based Review Sentence Selection for Opinion Targets Extraction and Opinion Orientation: A Distant Supervision Approach”, *International Journal of Intelligent Engineering and Systems*, Vol. 11, No. 5, pp. 114–124, 2018.
- [18] I. Cruz, A. F. Gelbukh, and G. Sidorov, “Implicit Aspect Indicator Extraction for Aspect based Opinion Mining”, *International Journal of Computational Linguistics and Applications*, Vol. 5, No. 2, pp. 135–152, 2014.
- [19] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, “SemEval-2015 Task 12: AspectBased Sentiment Analysis,” In: *Proc. of the 9th International Workshop on Semantic Evaluation*, p. 10, 2015.
- [20] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to Platt’s SMO Algorithm for SVM Classifier Design”, *Neural Computation*, Vol.13, pp.637-649, 2001.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update”, *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, p. 10, 2009.