Unsupervised Twitter Sentiment Classification

Mihaela Dinsoreanu and Andrei Bacu
Computer Science Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Keywords: Sentiment Classification, Unsupervised Learning, NLP, Word Sense Disambiguation.

Abstract: Sentiment classification is not a new topic but data sources having different characteristics require customized methods to exploit the hidden existing semantic while minimizing the noise and irrelevant information. Twitter represents a huge pool of data having specific features. We propose therefore an unsupervised, domain-independent approach, for sentiment classification on Twitter. The proposed approach integrates NLP techniques, Word Sense Disambiguation and unsupervised rule-based classification. The method is able to differentiate between positive, negative, and objective (neutral) polarities for every word, given the context in which it occurs. Finally, the overall tweet polarity decision is taken by our proposed rule-based classifier. We performed a comparative evaluation of our method on four public datasets specialized for this task and the experimental results obtained are very good compared to other state-of-the-art methods, considering that our classifier does not use any training corpus.

1 INTRODUCTION

Users now share in real time comments and opinions about companies, celebrities, events, and products through different social media applications. Sentiment classification has now become the dominant approach used for understanding emotions from online data and involves identifying the target and the polarity of opinions in unstructured text such as blogs, reviews, tweets, messages, and comments.

The classification results are defined categories such as positive, negative or neutral. However, in some cases the expressed emotions are multiple and can be mixed or simply irrelevant, given a specific topic. In such cases, the sentiment and meaning of the different words and phrases, as well as the main topic, are crucial in determining accurately the overall text polarity. Commonly, many words have a specific polarity by themselves but, in some cases the polarity of a word depends on the context in which the word is used. Hence, sentiment classification systems should include word sense disambiguation of each word according to the context. Traditional Natural Language Processing (NLP) techniques such as Part-of-Speech (POS) tagging, negation detection and topic recognition that are applicable for correctly written documents need to be enhanced and adapted for user generated content such as tweets.

For Twitter Sentiment Classification, the most common and challenging issues to solve are related to handling abbreviations, character repetitions, emoticons, misspelled words, and slang. In this paper we propose an unsupervised, domain independent, approach that aims for a sentiment classification method that addresses the aforementioned challenges of tweets while trying also to exploit the useful information specific to this type of data. The classifier works at the tweet (sentence) level of context.

The rest of the paper is structured as follows:

Section 2 discusses related research work that has been considered in our paper. Section 3 presents in detail our approach for Sentiment Classification on Twitter. Section 4 demonstrates the applicability of our system through a comparative evaluation on four evaluation datasets specialized for this task and presents the experimental results. Finally, Section 5 concludes the paper.

2 RELATED WORK

The problem of classifying sentiment polarity has received considerable attention from research communities, several approaches surveyed in (Pang and Lee, 2008). (Riloff, 2003; Whitelaw et al., 2005) consider lexical resources for identifying sentiment

Dinsoreanu M. and Bacu A..
Unsupervised Twitter Sentiment Classification.
DOI: 10.5220/0005079002200227
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2014), pages 220-227
ISBN: 978-989-758-050-5

words with positive and negative polarity, such as SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006) or QWordNet (Agerri and García-Serrano, 2010). Enhancements for these approaches include negation detection (Das and Chen, 2001; Wiegand et al., 2010) or the identification of words that boost the sentiment score of other words (Turney, 2002; Thelwall et al., 2010). Other relevant features that proved to be reasonably effective include emoticons (Derks, et al., 2008; Fullwood and Martino, 2007) and word abbreviations (Thurlow, 2003). However, sentiment words can have different meanings in different contexts (Esuli and Sebastiani, 2006; Andreevskaia and Bergler, 2006; Wiebe and Mihalcea, 2006), thus Word Sense Disambiguation has been used to improve sentiment classification (Akkaya et al., 2009). Other works simultaneously extract positive and negative sentiments from short informal text (Thelwall et al., 2010). Resources such as lookup tables for scores of sentiment words, negation words, emoticons or slang have been successfully used by (Thelwall et al., 2010; Thelwall et.al, 2012). A novel approach (Bravo-Marquez et al., 2013) combines aspects such as opinion strength, emotion and polarity indicators, generated by existing sentiment analysis methods and resources, and shows significant improvement in Twitter Sentiment Classification tasks such as polarity and subjectivity identification. Most of the work in the field addresses English documents. Traditional multilingual approaches rely either on translations from the source language to English (Denecke, 2008), or on cross-lingual training that requires additional resources such as parallel corpora in the source language and English (Mihalcea et al., 2007). More recent approaches targeting tweets (Narr et al., 2012) address the multilingual problem by a supervised solution that involves training classifiers for 4 languages and needing a set of raw tweets for training for any other considered language. Even so, authors report that classification performance in terms of accuracy varies significantly between languages.

To evaluate existing solutions, the Semantic Evaluation (SemEval) community has organized a workshop having as topic Sentiment Analysis in Twitter (Wilson et al., 2013). Results show that the strongest team (Mohammad et al., 2013), achieved a F1-measure of 69% on subtask B – Twitter that is of interest in this paper. Only one approach used an unsupervised method (Ortega et al., 2013), achieving a F1-measure of 50% on subtask B – Twitter. The rest of the participant systems were semi or fully supervised strategies. Finally, this year was proposed a rerun of SemEval-2013 task 2 (Nakov et al, 2013) as SemEval-2014 Sentiment Analysis Task 9[1], with new test data from Twitter and another genre.

3 CONCEPTUAL SOLUTION

Our approach is based on an unsupervised strategy consisting of three major NLP phases: POS-tagging, text preprocessing and contextual Word Sense Disambiguation and a final tweet polarity classification phase

Firstly, we employ a Twitter-aware POS-tagger and tokenizer in combination with an extensive preprocessing task on the input tweets. The resulted <word, POS, sense> triples are fed into a Word Sense Disambiguation (WSD) method based on a best-match and high-confidence score with the SentiWordNet database. Each relevant word is assigned a positive, negative, or objective (neutral) polarity, depending on the context in which it occurs. Finally, we propose a classification model in terms of a set of classification rules based on which the overall tweet polarity decision is taken. The conceptual architecture of our Twitter Sentiment Classifier is shown in Figure 1 below.

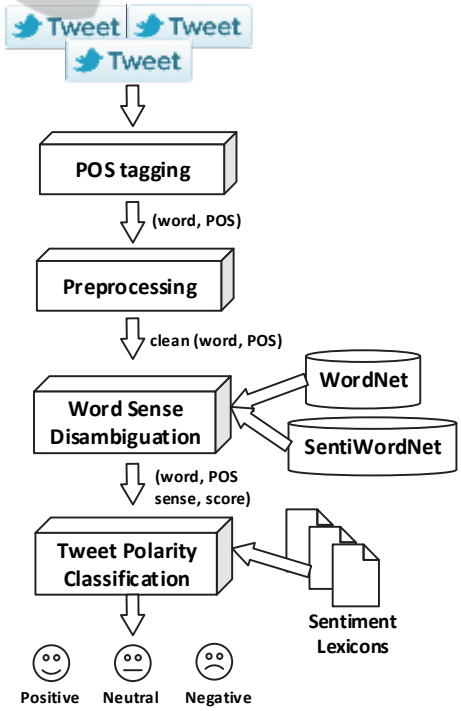

Figure 1: Conceptual System Architecture.

[1] http://alt.qcri.org/semeval2014/task9/

Being an unsupervised approach, we do not need any training data but our results depend on the existence of some lexical resources, namely the Twitter-aware POS tagger, sentiment lexicons, spelling corrector and the Word Sense Disambiguation method. Since our classifier does not rely on any training corpus and the proposed set of rules is general, we claim that our system is a domain-independent one. The approach is used on English tweets but it can be adapted to other languages given appropriate resources.

3.1 Lexical Resources for Sentiment Classification

Research communities have developed several lexical resources for sentiment classification. Firstly, a list of English words was annotated by (Wilson et al., 2005) with positive and negative sentiment categories, thus creating the Opinion Finder lexicon. The Affective Norms for English Words (ANEW) lexicon was released by (Bradley and Lang, 2009) and further enhanced by (Nielsen, 2011), leveraging the AFINN lexicon. The popular Wordnet lexical database introduced by (Miller et al, 1995) was extended by (Esuli and Sebastiani, 2006) through the addition of sentiment scores to synsets, creating SentiWordnet, updated to SentiWordNet 3.0 by (Baccianella et al., 2010). Finally, SentiStrength was leveraged by (Thelwall et.al, 2010) for estimating the sentiment strength and further improved by (Thelwall et.al, 2012) as SentiStrength 2. (Bravo-Marquez et al., 2013) have shown that the Twitter Sentiment Classification task is boosted by using different sentiment features.

In our approach, we make use of the sentiment scores attached to each synset of WordNet. However, not all words and expressions encountered in tweets are present in WordNet, even if spelling correction and lemmatization are applied. To handle this case, we combined the Emotion Lookup Table used in the SentiStrength 2 approach with the affective lexicon introduced by (Nielsen, 2011) and the Opinion Finder lexicon.

Moreover, we adapted and included several important features from the SentiStrength 2 approach (Thelwall et.al, 2012). Firstly, we refined the Booster Word List and used it to strengthen or weaken the score of following sentiment words. Secondly, two or more repeated letters and capitalization added to sentiment words, and repeated punctuation will give a score boost of +/- 1. Finally, two or more consecutive positive/negative terms having sentiment scores of at least +/- 3 increase the strength of the second word by +/- 1.

3.2 POS-tagging Phase

For the POS-tagging operation we have used the last version of a fast and robust Twitter-aware tokenizer and part-of-speech tagger (O'Connor B. et.al, 2013) for each input tweet. Besides using an extended tagset specialized for tagging tweets, it helps significantly in determining abbreviations, emoticons, hashtags, slang, and incorrect words. From the tests that we have done for our approach, slang words and expressions are tagged either as "!" (Interjection) or "G" (other abbreviations, foreign words, symbols, garbage).

The output obtained from this first phase is represented by <word, POS> pairs that will be submitted to an extensive cleaning and standardization process in the next phase.

3.3 Preprocessing Phase

Unlike text present in books or articles, tweets are limited to 140 characters. Given that, Twitter users include additional information with strong semantics such as abbreviations, emoticons, hashtags, slang or URLs. Therefore, text preprocessing is a requirement needed in order to eliminate noisy or recover, if possible, incomplete information.

First of all, we remove URLs, re-tweets and user mentions. Stopwords are eliminated by using the Natural Language Tool Kit[2] (NLTK) Stopwords Corpus. Tokens containing "#" (hashtags), frequently represent an emotion, thought or opinion regarding the tweet's topic, so we remove only the "#". Misspells are brought to a grammatical form by using a spelling correction algorithm[3]. Further, we developed a normalization module to delete repeated letters in a word in order to create a correct English word. For example: "*amaaaziing"* is converted into the correct form "*amazing"*.

Commonly used phrases (e.g. "*ain't")* are replaced with their grammatical form ("*is not"*) by making use of regular expressions. For this, we created a dictionary with the most commonly used idioms on Twitter and added an associated sentiment score (e.g. "*can't wait*": 3). Moreover, we leveraged two additional resources: an emoticons dictionary obtained from Wikipedia[4] and an emoticon website[5],

[2] http://www.nltk.org/

[3] http://norvig.com/spell-correct.html

[4] http://en.wikipedia.org/wiki/List_of_emoticons

and the NoSlang[6] online dictionary. Each emoticon, slang and abbreviation was manually annotated with a sentiment score. For example, positive emoticons such as ":-)", ":D" are annotated with sentiment scores of 1 and 2, negative emoticons such as ":-(", ":((" are annotated with sentiment scores of -1 and -2, and words such as *"thx", "h8"* are annotated with sentiment scores of 1 and -2. The range of the sentiment scores are from -5 (negative) to 5 (positive), just like in the AFINN and SentiStrength 2 approaches.

Finally, the preprocessed <word, POS> pairs that are obtained represent the input for the Word Sense Disambiguation (WSD) phase. The WSD algorithm considers only the pairs that have a valid POS-tag i.e. it is present in the SentiWordNet database: "A", "N", "R", and "V".

3.4 Word Sense Disambiguation

Many approaches have tried to determine the polarity of opinion using annotated lexicons with prior polarity (Kamps and Marx, 2002; Turney, 2002; Riloff, 2003; Whitelaw et al., 2005). However a word can modify the prior polarity in relation to the context within which it is invoked. For example the word "sick" is used with a negative meaning in the sentence: "I feel very sick today". Whereas it is used with a positive meaning in the sentence: "Your new laptop is too sick".

The chosen Word Sense Disambiguation method is the one proposed in the WordNet SenseRelate[7] project. The algorithm makes use of measures of semantic similarity and relatedness to obtain the contextual polarity of all words in tweets. Practically, the algorithm assigns to a word the meaning that is most related to a given set of words. We considered the following works on semantic similarity for our experiments: (Jiang and Conrath, 1997), (Banerjee and Pedersen, 2003) and (Patwardhan, 2003). Since the approach in (Jiang and Conrath, 1997) is limited to noun-noun concept pairs, we based our solution on (Patwardhan, 2003) that proved to have a higher WSD accuracy for our classification experiments.

The goal of the WSD process consists practically in determining the best <word, POS-tag, sense> match for each of the <word, POS-tag> pairs received as input from the previous phase. Lemmatization is also employed for a better matching percentage. The required context for each word sense is given by all the other <word, POS-tag> pairs belonging to the same tweet.

The first <WSD_word, WSD_POS, sense> triple i.e. the one with the highest confidence score, is considered a best match, given a <word, POS> pair, if word = WSD_word and POS = WSD_POS. Further, positive and negative sentiment scores are extracted from SentiWordNet, based on <word, POS, sense> matching. The obtained information is represented as follows: <word, POS, sense, PosScore, NegScore>.

Considering the work of (Pang et al., 2002), we introduced a negation detection method for every part of a tweet that starts with a negation word (e.g., don't, wouldn't) and ends with one of the punctuation marks: '.', ',', ':', ';', '!', '?' or any combination and repetition between them. The "_NEG" suffix was added to each word that follows the negation word. The list of negation words was adopted from Christopher Potts' sentiment tutorial (14). Let us consider an example tweet: *Dear @Apple, I don't want the newsstand icon on my screen. #notcool.* By applying the negation detection method, the obtained output is the following: *Dear @Apple, I don't want_NEG the newsstand_NEG icon_NEG on my screen_NEG. #notcool.*

Finally, we append a NEG flag to the existing tuple <word, POS, sense, PosScore, NegScore>, that inverts the polarity of sentiment words contained within it. This represents the end result of this phase.

3.5 Tweet Polarity Classification

In (Saif H. et.al, 2013) are presented in detail eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Tweets in these datasets have been annotated with different sentiment labels including: Negative, Neutral, Positive, Mixed, Other and Irrelevant. Our approach considers only positive, negative and neutral polarities. We will present in more detail all the evaluation datasets, along with the individual tweet polarity statistics in the next section.

In this phase we determine the overall tweet polarity by leveraging a rule-based classifier.
Consider the following notations:

w – token from tweet having sentiment/emotion score

score(w, p) – positive sentiment score of token w

score(w, n) – negative sentiment score of token w

score(w, o) – objective (neutral) sentiment score of token w

[5] http://cool-smileys.com/text-emoticons

[6] http://www.noslang.com/dictionary/

[7] http://senserelate.sourceforge.net

$\boldsymbol{Pos} = \sum_{w \in tweet} score(w,p)$
if (score(w,p)> 0 and score(w,p) > score(w,n))

$\boldsymbol{Neg} = \sum_{w \in tweet} score(w,n)$
if (score(w,n)> 0 and score(w,n) > score(w,p))

$\boldsymbol{Obj} = \sum_{w \in tweet} score(w,o)$
if (score(w,o) > 0 and score(w,p)=0 and score(w,n)=0)

$\boldsymbol{PosCnt} = \sum_{w \in tweet} w$
if (score(w,p)> 0 and score(w,p) > score(w,n))

$\boldsymbol{NegCnt} = \sum_{w \in tweet} w$
if (score(w,n)> 0 and score(w,n) > score(w,p))

$\boldsymbol{ObjCnt} = \sum_{w \in tweet} w$
if (score(w,o) > 0 and score(w,p)=0 and score(w,n)=0)

$\boldsymbol{TokCnt} = \sum_{w \in tweet} w$
if (score(w,p)>0 or score(w,n)>0 or score(w,o)> 0)

$\boldsymbol{PR} = \frac{\mathbf{PosCnt}}{\mathbf{TokCnt}}$ (positive count ratio)

$\boldsymbol{NR} = \frac{\mathbf{NegCnt}}{\mathbf{TokCnt}}$ (negative count ratio)

$\boldsymbol{OR} = \frac{\mathbf{ObjCnt}}{\mathbf{TokCnt}}$ (objective/neutral count ratio)

Based on the formulas presented in Table 1, the classifier determines the dominant sentiment expressed in the tweets i.e. positive, negative or neutral. The tweets that cannot be classified in one of the above mentioned polarity classes will be considered as containing mixed sentiment polarities and included in the neutral class.

Table 1: Classification conditions for each polarity class: positive, negative and neutral.

Polarity	Classification condition
positive	$\frac{\mathbf{Pos}}{\mathbf{Neg}} \geq \frac{3}{2}, \frac{\mathbf{Pos}}{\mathbf{Obj}} \geq \frac{3}{2}, \boldsymbol{PR > NR, PR > OR}$
negative	$\frac{\mathbf{Neg}}{\mathbf{Pos}} \geq \frac{3}{2}, \frac{\mathbf{Neg}}{\mathbf{Obj}} \geq \frac{3}{2}, \boldsymbol{NR > PR, NR > OR}$
neutral	$\frac{\mathbf{Obj}}{\mathbf{Pos}} \geq \frac{3}{2}, \frac{\mathbf{Obj}}{\mathbf{Neg}} \geq \frac{3}{2}, \boldsymbol{OR > PR, OR > NR}$

4 EXPERIMENTAL RESULTS

4.1 Comparison of Lexical Resources

We analyzed the overlapping words between all the lexical resources used in our approach: AFINN, OPFIND, SS2 and SWN3 and we observed that SWN3 is much larger than the other resources. However, the SWN3 database includes many objectve words (Pos = Neg = 0) that are not useful when assigning positive or negative polarity to tweets. Our rule-based classifier takes into account the objective score from SWN3 only if the respective word does not exist in any of the other lexical resources. For this particular case, we have managed to diminish the misclassification percentage for the positive and negative polarity classes by reducing the objective score from SWN3 with a factor of 5. This decision is strongly based on the tests done and experimental results obtained for each polarity class.

Analyzing further the AFINN lexicon, it can be seen that it contains some words that are not included in none of the other resources. Moreover, by comparing several words and the annotated sentiment values from each of the lexical resources used,we noticed that there exists some support among different resources: some words that have positive and negative sentiment scores in AFINN, SS2 and SWN3 have the same category in OpinionFinder. However, other words such as “thanks” and “sympathy” can have different sentiment values in SWN3 or simply do not exist – “sympathy” in SS2. These words can be used to express either positive or negative opinions, depending on the context, issue approached in the WSD phase.

4.2 Evaluation Datasets

We considered four evaluation datasets for our experiments: Stanford Twitter Sentiment Test Set (STS-Test), STS-Gold, Sanders Twitter Dataset and the SemEval-2013 Dataset (SemEval).

The test set (STS-Test) of the Stanford Twitter sentiment corpus[8] was introduced by (Go et al., 2009) The STS-Gold dataset was introduced in (Saif et al., 2013) The Sanders Twitter Dataset[9] consists of 5,512 tweets on four different topics (Apple, Google, Microsoft, and Twitter). The tweets are oriented mostly on product reviews, leading to domain dependence. The SemEval dataset was constructed for the Twitter Sentiment Analysis Task 2 (Nakov et al., 2013) and used in the SemEval-2013[10] and SemEval-2014 Sentiment Analysis Task 9[11] competitions. We have used the tweet ids provided by (Nakov et al., 2013) and managed to

[8] http://help.sentiment140.com/
[9] http://www.sananalytics.com/lab/twitter-sentiment/
[10] http://www.cs.york.ac.uk/semeval-2013/task2/
[11] http://alt.qcri.org/semeval2014/task9/

download 8,696 tweets with 3,631 negative, 1,405 neutral and 3,660 positive tweets.

For all the evaluation datasets, each tweet was annotated with a positive, negative or neutral polarity. Table 2 shows the statistics for all four evaluation datasets. It can be observed that the most balanced evaluation datasets are STS-Test and SemEval. On the other hand, the measures obtained for the STS-Gold and Sanders datasets will be affected by the imbalance between the tweets from the three polarity classes, as it can be seen below.

Table 2: Statistics of Evaluation Datasets.

	STS-Test	STS-Gold	Sanders	SemEval
#negative	177	1,402	654	3631
#neutral	139	77	2,503	1405
#positive	182	632	570	3660
#total	498	2,205	3,727	8696

4.3 Sentiment Classification Results

We performed the Sentiment Classification task using our rule-based classifier on four datasets: STS-Test, STS-Gold, Sanders and SemEval. For each dataset we selected only the subset of positive, negative and neutral tweets. We considered as performance measures accuracy, precision, recall and the F-measure for each polarity class. In tables 3, 4 and 5 we present the classification results on the mentioned evaluation datasets, for each polarity class: positive, negative and neutral.

Table 3: Accuracy, precision, recall and F-measure for the positive polarity class.

	STS-Test	STS-Gold	Sanders	SemEval
Accuracy	81.908	**83.764**	75.261	78.737
Precision	**81.818**	76.144	54.074	80.475
Recall	**91.443**	86.867	86.140	84.907
F-positive	**86.363**	81.152	66.441	82.632

For the positive class, the highest measures are achieved on the STS-Test dataset: precision = 81.818%, recall = 91.443% and F-positive = 86.363%. It is also worth noticing that the per-class performance is highly affected by the distribution of positive, negative and neutral tweets in the dataset. Moreover, by comparing the positive class and negative class measures on the SemEval dataset, we can see a clear performance gain in classification towards the first class. Taking into account the evaluation results and the fact that the most balanced datasets are STS-Test and SemEval, we conclude that our approach is better at detecting positive tweets than detecting negative tweets.

For the negative class, the highest measures are achieved on the STS-Gold dataset: precision = 94.098%, recall = 83.024% and F-negative = 88.215%.

Table 4: Accuracy, precision, recall and F-measure for the negative polarity class.

	STS-Test	STS-Gold	Sanders	SemEval
Accuracy	81.908	**83.764**	75.261	78.737
Precision	85.628	**94.098**	57.305	62.056
Recall	80.790	**83.024**	76.758	73.451
F-negative	83.139	**88.215**	65.620	67.275

However, the number of negative tweets from this dataset is approximately double as compared to the sum between the positive and neutral tweets. Comparing with the evaluation results from the positive class on the STS-Test, STS-Gold and SemEval datasets, it can be observed that the measures from for the negative class are clearly influenced by the imbalanced number of tweets in each polarity class.

Table 5: Accuracy, precision, recall and F-measure for the neutral polarity class.

	STS-Test	STS-Gold	Sanders	SemEval
Accuracy	81.908	**83.764**	75.261	78.737
Precision	77.165	54.251	**93.257**	85.321
Recall	70.503	**78.362**	72.393	74.644
F-positive	73.684	64.114	**81.511**	79.626

For the neutral class, the highest measures are achieved on the Sanders and STS-Gold datasets: precision = 93.257%, recall = 78.362% and F-neutral = 81.511%. Again, notice a performance gain for the Sanders dataset that contains much more neutral tweets than positive and negative. On the rest of the evaluation datasets the results are somewhat modest as compared to the positive and negative classes. However, considering the difficulty of correctly classifying tweets having neutral (objective) polarity, we conclude that our results are good, as compared with the other polarity classes.

As was also done in (Liu et al., 2012; Bravo-Marquez et al., 2013), we focused more on the accuracy and F1 measure than on precision and recall, because both accuracy and the F1 measure are affected by both false positive and false negative classification results.

In Table 6 below we included also the average accuracy, precision, recall and F-measure. The highest accuracy and recall is achieved on the STS-Gold dataset, with 83.764% and 82.751%

respectively. For the other two measures, the highest precision and F1-score is achieved on the STS-Test dataset, with 81.537% and 81.062% respectively.

Table 6: Average measures: accuracy, precision, recall and the harmonic mean (F measure).

	STS-Test	STS-Gold	Sanders	SemEval
Accuracy	81.908	**83.764**	75.261	78.737
Precision	**81.537**	74.831	68.212	75.950
Recall	80.912	**82.751**	78.430	77.667
F1-score	**81.062**	77.827	71.190	76.511

We compare our work with (Bravo-Marquez et al., 2013) on the STS-Test and Sanders datasets, with (Mohammad et al., 2013) on the SemEval dataset. Furthermore, we compared our results on all the four evaluation datasets with the results presented by (Saif et al., 2013).

Firstly, our classifier outperforms the accuracy precision, recall and F1-score of the Baselines used by (Bravo-Marquez et al., 2013) for the STS-Test dataset by roughly 4%. On the Sanders dataset, we found the results not so conclusive, especially because of the distribution of positive, negative and neutral tweets.

Secondly, our F-score of 76.51% outperformed the F-score of 69.02% reported by (Mohammad et al., 2013) on the SemEval (2013) dataset. We consider this to be the most important performance indicator of our rule-based classifier.

Finally, we compare our work with (Saif et al., 2013), which performed only a binary sentiment classification using a MaxEnt classifier, so the neutral class was not considered in the results. The accuracy and the average F-measure reported for our classifier is slightly better on the STS-Test dataset. On the STS-Gold dataset, we obtained a lower accuracy by 2%, but a higher average F-measure by 2%. Again, the results we report on the Sanders dataset are not so good, given the high number of neutral tweets. Last, we conclude that the results on the SemEval dataset are good – only a 4% difference of the evaluation measures.

5 CONCLUSIONS

In this paper, we have presented a novel, unsupervised, approach for Twitter Sentiment unsupervised approach for Twitter Sentiment Classification based on our rule-based classifier. The NLP pipeline combines several sentiment analysis methods and uses some existing lexical resources. More specific, our system relies on an unsupervised strategy that uses a Twitter-aware POS-tagger and tokenizer in combination with an extensive preprocessing task to produce input for a Word Sense Disambiguation (WSD) method. The method is able to differentiate between positive, negative, and objective (neutral) polarities for every word, given the context in which it occurs. Based on the rule-based classification model we propose, the overall tweet polarity decision is taken. The experimental results prove that our proposal is accurate for this complex task, given that our approach does not use any training corpus.

As future work we aim to consider and include other lexical resources and sentiment analysis methods that can improve the current system. We plan to evaluate our approach on all the datasets surveyed by (Saif et al., 2013), to compare them with other similar works and improve our classifier. Also, we want to evaluate our approach on the datasets used in SemEval-2014 task 9.

REFERENCES

Agerri, R., García-Serrano, A. (2010, May). Q-WordNet: Extracting Polarity from WordNet Senses. In LREC.

Akkaya, C., Wiebe, J., Mihalcea, R. (2009, August). Subjectivity word sense disambiguation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 190-199). Association for Computational Linguistics.

Andreevskaia, A., Bergler, S. (2006, April). Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In EACL (Vol. 6, pp. 209-215).

Baccianella, S., Esuli, A., Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

Banerjee, S., Pedersen, T. (2003, August). Extended gloss overlaps as a measure of semantic relatedness. In IJCAI (Vol. 3, pp. 805-810).

Bradley, M. M., Lang, P. J. Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings. Technical Report C-1, The Center for Research in Psychophysiology University of Florida, 2009.

Bravo-Marquez, F., Mendoza, M., Poblete, B. (2013, August). Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (p. 2). ACM.

Das, S., Chen, M., (2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of APFA-2001. 2001.

Denecke, K. (2008, April). Using Sentiwordnet for multilingual sentiment analysis. In Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on (pp. 507-512). IEEE.

Esuli, A., Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC (Vol. 6, pp. 417-422).

Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

Jiang, J. J., Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Kamps, J., Marx M., (2002). Words with attitude. In Proceedings of the 1st International Conference on Global WordNet, pages 332–341. CIIL, Mysore India.

Liu, K. L., Li, W. J., Guo, M. (2012). Emoticon Smoothed Language Models for Twitter Sentiment Analysis. In AAAI.

Mihalcea, R., Banea, C., Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In ANNUAL MEETING-ACL (Vol. 45, No. 1, p. 976).

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Mohammad, S. M., Kiritchenko, S., Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter.

Narr, S., Hülfenhaus, M., Albayrak, S. (2012). Language-independent Twitter sentiment analysis. In KDML workshop on knowledge discovery, data mining and machine learning.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

Ortega R., Fonseca A., Gutierrez Y. and Montoyo A.,(2013) SSA-UO: Unsupervised Twitter Sentiment Analysis, in SemEval 2013.

Pang B., Lee L., and Vaithyanathan S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceeding of Empirical Methods in Natural Language Processing, pages 79–86.

Pang B, Lee L. (2008) Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1-2, pp. 1-135, 2008.

Patwardhan S., (2003). Incorporating dictionary and corpus information into a Context Vector Measure of Semantic Relatedness. Master's thesis, Dept. of Computer Science, University of Minnesota, Duluth.

Riloff, E., Wiebe, J. (2003) Learning Extraction Patterns for Subjective Expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03).

Saif, H., Fernandez, M., He, Y., Alani, H. (2013). Evaluation datasets for twitter sentiment analysis. In Proceedings ESSEM in Conjunction with AI* IA Conference, Turin, Italy.

Thelwall, M., Buckley, K., Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163-173.

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. Cyberemotions, 1-14.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Thurlow, C., & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. Discourse analysis online, 1(1), 30.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Pages 417–424.

Wiebe, J., Mihalcea, R. (2006). Word sense and subjectivity. In Proceedings of the 21st ICCL and the 44th annual meeting of the ACL (pp. 1065-1072).

Wiegand M., Balahur A., Roth B., Klakow D., Montoyo A. (2010). A survey on the role of negation in sentiment analysis. In Proceedings of NeSp-NLP '10, pages 60–68.

Wilson T., Kozareva Z., Nakov P., Rosenthal S., Stoyanov V., Ritter A. (2013). SemEval-2013 task 2: Sentiment analysis in twitter. In Proceedings of the International Workshop on Semantic Evaluation, SemEval 2013, June.

Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354).

Wilson, T., Wiebe, J., Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 35(3), 399-433.