BUAP: Polarity Classification of Short Texts

David Pinto[1], Darnes Vilariño[1], Saul León[1], Miguel Jasso[1,2], Cupertino Lucero[2]

[1] Benemérita Universidad Autónoma de Puebla
14 Sur y Av. San Claudio, CU, 72570
Puebla, Puebla, México
{dpinto,darnes,saul.leon}@cs.buap.mx

[2] Universidad Tecnológica de Izúcar de Matamoros
Prolongación Reforma 168, Santiago Mihuacan, 74420
Izúcar de Matamoros, Puebla, México
migueljhdz18@yahoo.com.mx, cuper_lucero@hotmail.com

Abstract

We report the results we obtained at the subtask B (Message Polarity Classification) of SemEval 2014 Task 9. The features used for representing the messages were basically trigrams of characters, trigrams of PoS and a number of words selected by means of a graph mining tool. Our approach performed slightly below the overall average, except when a corpus of tweets with sarcasm was evaluated, in which we performed quite well obtaining around 6% above the overall average.

1 Introduction

Analyzing polarity in texts is an important task that may have various applications in real life. There exist plenty of tasks that may be benefited of computational procedures that automatically allow to detect if the author intention has been to express himself as a positive, negative, neutral or objective manner. Let us consider, for instance, when a public figure (such as a politician, celebrity, or business leader) would like to investigate its reputation in public media. Another example would be to calculate the reputation of a public or private institution. In any case, the construction of methods for determining the polarity of messages at Internet would help to investigate their reputation.

In this paper, we present the results we obtained when we carried out experiments for the subtask B

of Semeval 2014 Task 9, which was named "Message Polarity Classification", and was defined as follow: "Given a message, decide whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen".

The remaining of this paper is structured as follows. In Section 2 we present some related work found at the literature with respect to the identification of emotions in short texts such as twitter. Section 3 presents the description of the features and classification model used in our experiments. The results obtained together with a discussion of these results are given in Section 4. Finally, the conclusions are given in Section 5.

2 Related Work

There exist a number of works in literature associated to the automatic identification of emotions in Twitter, mainly due to the massification of this social network around the world and the easy manner we can access to the Tweets from API's provided by Twitter itself. Some of these works have focused on the contribution of some particular features, such as Part of Speech (PoS) tags, emoticons, etc. on the aforementioned task. In Agarwal et al. (2011), for example, the a priori likelihood of each PoS is calculated. They use up to 100 additional features that include emoticons and a dictionary of positive and negative words. They have reported a 60% of accuracy in the task. On the other hand, in Mukherjee and Bhattacharyya (2012), a strategy based on discursive relations, such as conectiveness and con-

ditionals, with low number of lexical resources is proposed. These relations are integrated in classical models of representation like bag of words with the aim of improving the accuracy values obtained in the process of classification. The influence of semantic operators such as modals and negations are analyzed, in particular, the degree in which they affect the emotion present in a given paragraph or sentence.

One of the major advances obtained in the task of sentiment analysis has been done in the framework of the SemEval competition. In 2013, several teams have participated with different approaches Becker et al. (2013); Han et al. (2013); Chawla et al. (2013); Balahur and Turchi (2013); Balage Filho and Pardo (2013); Moreira et al. (2013); Reckman et al. (2013); Tiantian et al. (2013); Marchand et al. (2013); Clark and Wicentwoski (2013); Hamdan et al. (2013); Martínez-Cámara et al. (2013); Levallois (2013). Most of these works have contributed in the mentioned task by proposing methods, techniques for representing and classifying documents towards the automatic classification of sentiment in Tweets.

3 Description of the Presented Approach

We have employed a supervised approach based on machine learning in which we construct a classification model using the following general features obtained from the training corpus.

1. Character trigrams

2. PoS tags trigrams

3. Significant Tweet words obtained by using a graph mining tool known as SubDue

The description of how we calculated each feature in order to construct a representation vector for each message is given as follows.

The probability of each character trigram given the polarity class, $P(trigram|class)$, was calculated in the training corpus. Thereafter, we assigned a normalized probability to each sentence polarity by combining the probability of each character trigram of the sentence, i.e., $\sum_{i=1}^{|message|} log\left[P(trigram_i|class)\right]$. Since we have four classes ("positive","negative","neutral"

and "objective"), we have obtained four features for the final vectorial representation of the message.

We then calculated other four features by performing a similar calculation than the previous one, but in this case, using the PoS tags of the message. For this purpose, we used the Twitter NLP and Part-of-Speech Tagging tool provided by the Carnegie Mellon University (Owoputi et al., 2013). Since the PoS tag given by this tool is basically a character, then the same procedure can be applied.

We performed preliminary experiments by using these eight features on a trial corpus, and we observed that the results may be improved by selecting significant words that may not be discovered by the statistical techniques used until now. So, we decided to make use of techniques based on graph mining for attempting to find those significant words. In order to find them, we constructed a graph representation for each message class ("positive","negative","neutral" and "objective"), using the training corpus. The manner we constructed those graphs is shown as follows.

Formally, given a graph $G = (V, E, L, f)$ with V being the non-empty set of vertices, $E \subseteq V \times V$ the edges, L the tag set, and $f : E \rightarrow L$, a function that assigns a tag to a pair of associated vertices. This graph-based representation attempt to capture the sequence among the sentence words, so as the sequence among their PoS tags with the aim of feeding a graph mining tool which may extract relevant features that may be further used for representing the texts. Thus, the set V is constructed from the different words and PoS of the target document.

In order to demonstrate the way we construct the graph for each short text, consider the following message: "ooh i love you for posting this :-)". The associated graph representation to this message is shown in Figure 1.

Once each paragraph is represented by means of a graph, we apply a data mining algorithm in order to find subgraphs from which we will be able to find the significant words which will be, in our case, basically, the nodes of these subgraphs. Subdue is a data mining tool widely used in structured domains. This tool has been used for discovering structured patterns in texts represented by means of graphs Olmos et al. (2005). Subdue uses an evaluation model named "Minimum encoding", a tech-
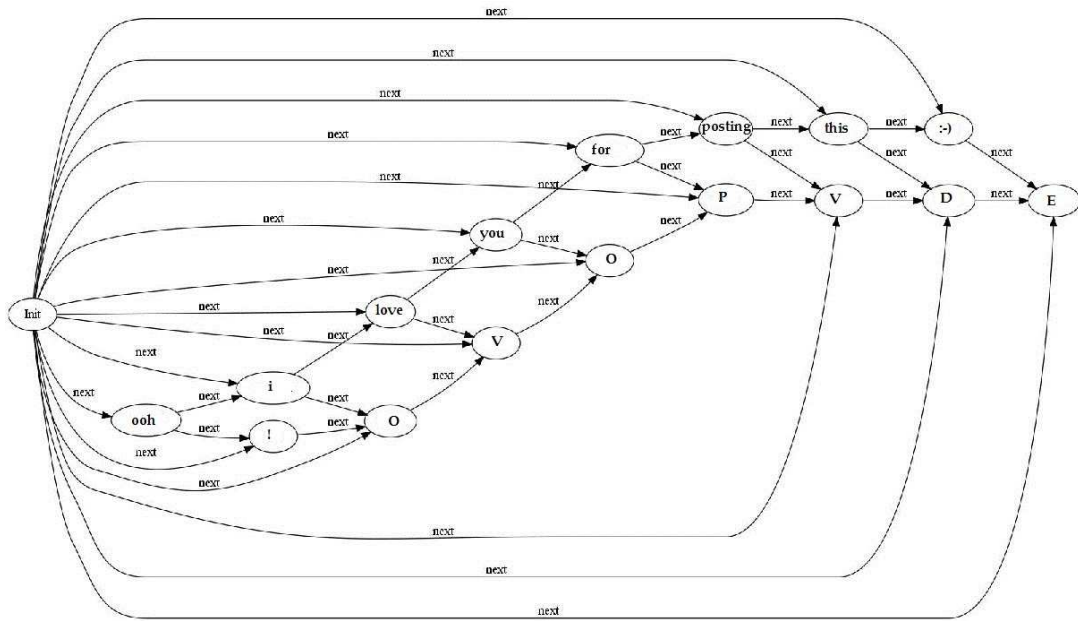
Figure 1: Graph based message representation with words and their corresponding PoS tags

nique derived from the minimum description length principle Rissanen (1989), in which t he best graph sub-structures are chosen. The best subgraphs are those that minimize the number of bits that represent the graph. In this case, the number of bits is calculated consi dering the size of the graph adjancency matrix. Thus, the best substructure is the one that minimizes $I(S) + I(G|S)$, where $I(S)$ is the number of bits required to describe the sub structure S, and $I(G|S)$ is the number of bits required to describe graph G after it has been compacted by the substructure S.

By applying this procedure we obtained 597 sigical negative words, 445 positive words, 616 objective words and 925 positive words. For the final representation vector we compiled the union of these words, obtaining 1915 significant words. Therefore, the total number of features for each message was 1,923.

We have used the training corpus provided at the competition (Rosenthal et al., 2014), however, we removed all those messsages tagged as the class "objective-OR-neutral", because all these messages introduced noise to the classification process. In total, we constructed 5,217 vectors of message representation which fed a support vector machine classi-

fier. We have used the SVM implementation of the WEKA tool with default parameters for our experiments (Hall et al., 2009). The obtained results are shown in the next section.

4   Experimental Results

The test corpus was made up short texts (messages) categorized as: "LiveJournal2014", "SMS2013", "Twitter2013", "Twitter2014" and "Twitter2014Sarcasm". A complete description of the training and test datasets can be found at the task description paper (Rosenthal et al., 2014).

In Table 1 we can see the results obtained at the competition. Our approach performed in almost all the cases slightly below to the overall average, except when we processed the corpus of Twitter with Sarcasm characteristics. We consider that two main problems were the cause of this result: 1) The corpus was very unbalanced and our approaches for alleviating this problem were not sufficient, and 2) From our particular point of view, there were a high difference between the vocabulary of the training and the test corpus, thus, leading the classification model to fail.

Table 1: Results obtained at the substask B of the Semeval 2014 Task 9

System	LiveJournal2014	SMS2013	Twitter2013	Twitter2014	Twitter2014Sarcasm	Average
NRC-Canada-B	74.84	70.28	70.75	69.85	58.16	68.78
CISUC_KIS-B-late	74.46	65.90	67.56	67.95	55.49	66.27
coooolll-B	72.90	67.68	70.40	70.14	46.66	65.56
TeamX-B	69.44	57.36	72.12	70.96	56.50	65.28
RTRGO-B	72.20	67.51	69.10	69.95	47.09	65.17
AUEB-B	70.75	64.32	63.92	66.38	56.16	64.31
SWISS-CHOCOLATE-B	73.25	66.43	64.81	67.54	49.46	64.30
SentiKLUE-B	73.99	67.40	69.06	67.02	43.36	64.17
TUGAS-B	69.79	62.77	65.64	69.00	52.87	64.01
SAIL-B	69.34	56.98	66.80	67.77	57.26	63.63
senti.ue-B	71.39	59.34	67.34	63.81	55.31	63.44
Synalp-Empathic-B	71.75	62.54	63.65	67.43	51.06	63.29
Lt_3-B	68.56	64.78	65.56	65.47	47.76	62.43
UKPDIPF-B	71.92	60.56	60.65	63.77	54.59	62.30
AMI_ERIC-B	65.32	60.29	70.09	66.55	48.19	62.09
ECNU-B	69.44	59.75	62.31	63.17	51.43	61.22
LyS-B	69.79	60.45	66.92	64.92	42.40	60.90
SU-FMI-B-late	68.24	61.67	60.96	63.62	48.34	60.57
NILC_USP-B-twitter	69.02	61.35	65.39	63.94	42.06	60.35
CMU-Qatar-B-late	65.63	62.95	65.11	65.53	40.52	59.95
columbia_nlp-B	68.79	59.84	64.60	65.42	40.02	59.73
CMUQ-Hybrid-B-late	65.14	61.75	63.22	62.71	40.95	58.75
Citius-B	62.40	57.69	62.53	61.92	41.00	57.11
KUNLPLab-B	63.77	55.89	58.12	61.72	44.60	56.82
USP_Biocom-B	67.80	53.57	58.05	59.21	43.56	56.44
UPV-ELiRF-B	64.11	55.36	63.97	59.33	37.46	56.05
Rapanakis-B	59.71	54.02	58.52	63.01	44.69	55.99
DejaVu-B	64.69	55.57	57.43	57.02	42.46	55.43
GPLSI-B	57.32	46.63	57.49	56.06	53.90	54.28
Indian_Inst_of_Tech-Patna-B	60.39	51.96	52.58	57.25	41.33	52.70
BUAP-B	**53.94**	**44.27**	**56.85**	**55.76**	**51.52**	**52.47**
SAP-RI-B	57.86	49.00	50.18	55.47	48.64	52.23
UMCC_DLSI_Sem	53.12	50.01	51.96	55.40	42.76	50.65
Alberta-B	52.38	49.05	53.85	52.06	40.40	49.55
SINAI-B	58.33	57.34	50.59	49.50	31.15	49.38
IBM_EG-B	59.24	46.62	54.51	52.26	34.14	49.35
SU-sentilab-B-tweet	55.11	49.60	50.17	49.52	31.49	47.18
lsis_lif-B	61.09	38.56	46.38	52.02	34.64	46.54
IITPatna-B	54.68	40.56	50.32	48.22	36.73	46.10
UMCC_DLSI_Graph-B	47.81	36.66	43.24	45.49	53.15	45.27
University-of-Warwick-B	39.60	29.50	39.17	45.56	39.77	38.72
DAEDALUS-B	40.83	40.86	36.57	33.03	28.96	36.05
Overall average	63.81	55.82	59.72	60.30	45.43	57.02

5 Conclusions

We have presented an approach for detecting message polarity using basically three kind of features: character trigrams, PoS tags trigrams and significant words obtained by means of a graph mining tool. The obtained results show that these features were not sufficient for detecting the correct polarity of a given message with high precision. We consider that the unbalanced characteristic and the fact the vocabulary changed significantly from the training to the test corpus influenced the results we obtained at the competition. However, a deep analysis we plan to do to the datasets evaluated will allow us in the fu-ture to find more accurate features for the message polarity detection task.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June 2011.

Pedro Balage Filho and Thiago Pardo. Nilc_usp: A hybrid system for sentiment analysis in twitter messages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2:*

Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 568–572, Atlanta, Georgia, USA, June 2013.

Alexandra Balahur and Marco Turchi. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 49–55, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA.

Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 333–340, Atlanta, Georgia, USA, June 2013.

Karan Chawla, Ankit Ramteke, and Pushpak Bhattacharyya. Iitb-sentiment-analysts: Participation in sentiment analysis in twitter semeval 2013 task. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 495–500, Atlanta, Georgia, USA, June 2013.

Sam Clark and Rich Wicentwoski. Swatcs: Combining simple classifiers with estimated accuracy. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 425–429, Atlanta, Georgia, USA, June 2013.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.

Hussam Hamdan, Frederic Béchet, and Patrice Bellot. Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (Se-*

mEval 2013), pages 455–459, Atlanta, Georgia, USA, June 2013.

Qi Han, Junfei Guo, and Hinrich Schuetze. Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 520–524, Atlanta, Georgia, USA, June 2013.

Clement Levallois. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 414–417, Atlanta, Georgia, USA, June 2013.

Morgane Marchand, Alexandru Ginsca, Romaric Besançon, and Olivier Mesnard. [lvic-limsi]: Using syntactic features and multi-polarity words for sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 418–424, Atlanta, Georgia, USA, June 2013.

Eugenio Martínez-Cámara, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia, and L. Alfonso Ureña López. Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 402–407, Atlanta, Georgia, USA, June 2013.

Silvio Moreira, João Filgueiras, Bruno Martins, Francisco Couto, and Mário J. Silva. Reaction: A naive machine learning approach for sentiment classification. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 490–494, Atlanta, Georgia, USA, June 2013.

Subhabrata Mukherjee and Pushpak Bhattacharyya.

Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of COLING 2012*, pages 1847–1864, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Ivan Olmos, Jesus A. Gonzalez, and Mauricio Osorio. Subgraph isomorphism detection using a code based representation. In *FLAIRS Conference*, pages 474–479, 2005.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390, 2013.

Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. teragram: Rule-based detection of sentiment phrases using sas sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 513–519, Atlanta, Georgia, USA, June 2013.

Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1989. ISBN 981020311X.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, 2014.

Zhu Tiantian, Zhang Fangxi, and Man Lan. Ecnucs: A surface information based system description of sentiment analysis in twitter in the semeval-2013 (task 2). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 408–413, Atlanta, Georgia, USA, June 2013.