Let's Agree to Disagree: Measuring Agreement between Annotators for Opinion Mining Task

Octavio Sánchez-Velázquez and Gerardo Sierra

Universidad Nacional Autónoma de México, Grupo de ingeniería lingüística, Instituto de Ingeniería UNAM,
oct_sanc@unam.mx, gsierram@iingen.unam.mx

Abstract. There is a need to know up to what degree humans can agree when classifying a sentence as carrying some sentiment orientation. However, a little research has been done on assessing the agreement between annotators for the different opinion mining tasks. In this work we present an assessment of agreement between two human annotators. The task was to manually classify newspaper sentences into one of three classes. For assessing the level of agreement, Cohen's kappa coefficient was computed. Results show that annotators agree more for negative classes than for positive or neutral. We observed that annotators might agree up to a level of substantial agreement of 0.65 for the best case or 0.30 for the worst.

Keywords: inter-annotator agreement, opinion mining, corpus labelling

1 Introduction

Nowadays there are more user-generated content than ever before. This allows people to quickly share and find all sort of information. This information is split in two big kinds: objective information, which is all the information that we can measure or prove, and subjective information which comprises appraisals, thoughts and opinions [20, 10]. The computational study of the latter is known as opinion mining. There has been several approaches to achieve this. The main two have been machine learning classifiers, which require tagged examples as inputs [14, 1], and lexicon based systems, which require lexicons with each word classified with its polarity (negative, positive, neutral) [6, 17, 12].

In order to create systems capable of detecting and classifying opinions written in texts, we need human labelled corpus, either to evaluate, generate the models or train a system. When extracting opinions from product reviews, there is usually a score readily available. This is because pages that let make a review of a product usually have a system for ranking the product being reviewed, usually with a system of stars. There are some domains, as newspaper articles, where the label for the document or the sentence are lacking. For these kind of texts there is a need of documents labelled by human annotators. While usually people might agree on marking objective tokens from the reality, when dealing with subjective information, such as opinions, there are several variables that

pp. 9–19; rec. 2016-02-09; acc. 2016-03-06

depend on the experiences of the annotator. Some of these might be the domain expertise, the personal opinion on some domain or subject, point of view with respect some idea, ambiguity of language or ambiguity on interpretation [2, 11, 4].

Even though this is known, just some papers report the agreement of the annotators. There are several works assessing inter-annotator agreement in different tasks, such as image annotation [13], part-of-speech tagging [3], word sense disambiguation [19]. There are also work done on other areas, as biology [7] or medicine [8]. As far as we know, there are just few works on opinion annotation agreement. These works are not as conclusive as they could be.

In this paper, we present a work on inter-annotator agreement assessment and we try to show that this agreement is lower than in some other areas. We used Cohen's kappa to assess the degree of agreement of two annotators. To explain this, in section 2 we review some work done in particular on opinion mining labelling agreement. Then on section 3, we briefly describe the Cohen's kappa (κ) that was the measure we used to assess the agreement. On section 4 we describe the data used for assessing this agreement and describe the experiment. We discuss the results of these experiment on section 5. Finally, in section 6 we present our conclusions and future work.

2 Related Work

In [16], the authors compare the agreement between experts and non experts. They use the Amazon Mechanical Turk (AMT) to perform five different tasks. These were affect recognition, word similarity, recognizing textual entailment, temporal event recognition, and word sense disambiguation. What the assess is how much did AMT agreed with a gold standard annotated by experts. For the task of affect recognition (which is a similar one to opinion mining), they used headlines of newspapers. To assess the agreement they used Pearson correlation. They report an agreement of 0.576 between expert annotators and 0.417 between a Non-expert and an expert. However when they compared the correlation between an expert and the average classification of all other experts and non-experts, they got 0.603. For the other tasks, their correlations were about 0.96. Which means that it is harder to agree with other people respect affect recognition.

In [5], the authors present an annotation scheme for adding entity and event target annotations to the MPQA corpus [18]. While they describe the added features to the MPQA, they have a section dedicated to the agreement. For the annotations, they developed a manual and gave it to the annotators. After the annotator were trained, they tagged four documents. To asses the agreement, they computed the F-measure of the documents taking turns of documents annotated by annotator A and those of annotator B as gold standard. By doing this, they report an agreement of 0.82. To solve the disagreement, annotators could discuss the situation so, after doing this, they could agree.

In [15] the authors present a method for creating a dictionary and marking the words with emotions or polarity. They annotated the words by multiple annotators. They used six basic emotions to annotate the words. They analysed the tags given by multiple annotators using a weighted kappa adapted for multiple entries. They trained the annotators with a manual that instructed how to label the words. They obtained an average of 0.6452 for a lineal weighted kappa and an average of 0.7717 for a quadratic weighted kappa. Differently from our work, they tagged just words and the frequency of each word used in the context that represents one of the six basic emotions. Based on this they propose a a factor of probability of affective use (FPA).

3 Cohen's kappa

Cohen's kappa is a statistic which assesses the inter-annotator agreement for categorical items. Opposite to a simple percent agreement computation, this measure takes into account that some agreement might take place by chance. It is thought to measure how much two annotators, when each classify N items into C mutually exclusive categories, agree with each other.

If we build a square matrix where each column represent the instances sorted in a category by one annotator and each row represents the instances sorted in the category by the other annotator, we could have a confusion matrix. In this matrix, we should have a correspondence in categories in each row and column. This means that the diagonal would represent the coincidences while classifying the instances.

In order to get the value of kappa (κ), it is needed to know the observed agreement. The observed agreement is how much, of the total, did the annotators agree. This is, what is the sum of the diagonal of our confusion matrix over the total of instances:

$$P(o) = \frac{\sum_{i=1}^{|C|} C_{ii}}{N}$$

where C is the confusion matrix containing the categories that were classified and in each cell the total of instances that were classified into that category for each annotator; and N is the total of instances that were classified.

However, it is possible that some of the coincidences between the two annotators is given by chance. In order to rule out this possibility, the hypothetical agreement by chance should be computed. This is done by:

$$P(e) = \sum_{i=1}^{|C|} \frac{C_{i+}}{N} \times \frac{C_{+i}}{N}$$

where C is the confusion matrix, C_{i+} is the sum of all the counts in row i and C_{+i} is the sum of all the counts in column i, and N is the total of instances that were classified.

Thus, in order to know the real agreement between the two annotators, it is needed to subtract the probability of chance agreement from the observed agreement. Once we take this into account we have the Conhen's kappa, given by:

$$\kappa = \frac{P(o) - P(e)}{1 - P(e)}$$

where $P(o)$ is the observed agreement and $P(e)$ is the hypothetical probability of chance agreement.

One aspect that has to be observed, is that this kappa assumes its theoretical maximum value of 1 only when both observers distribute codes the same. So, when it is the case that corresponding row and column are not identical, we need to compute the maximum possible value and use that to interpret in a better way the agreement assessed. This is done by the next formula:

$$\kappa_{max} = \frac{P_{max} - P(e)}{1 - P(e)}$$

with:

$$P(e) = \sum_{i=1}^{|C|} \frac{C_{i+}}{N} \times \frac{C_{+i}}{N}, P_{max} = \sum_{i=1}^{|C|} \min\left(\frac{C_{i+}}{N}, \frac{C_{+i}}{N}\right)$$

where $P(e)$ is the same as before and $Pmax$ is the sum of the minimum number between the row and the correspondent column over the total of instances classified N.

This help us to interpret the obtained kappa as a proportion of the theoretical maximum agreement that annotators could have.

Nevertheless, interpreting the value of kappa is a difficult task for different factors like the distribution of the probabilities of the categories, and the number of categories itself. However, there are some guidelines that has appeared and are widely used, particularly in the areas of health sciences and humanities. Arguably the most used are those proposed by Landis and Koch [9]. This is shown in the table 1

Table 1. Interpretation of kappa value

Kappa Statistic	*Strength of agreement*
<0.0	Poor
0.00 - 0.2	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost perfect

In this table, it is possible to see that according to the values of kappa, the strength of the agreement varies. We are going to use these guidelines and the maximum kappa in order to help in the interpretation of our values. We are going to use a division of the κ over the κ_{max} in order to obtain the proportion of agreement normalised by the maximum expected value.

4 Experimental set-up

In this section we explain the data we used to perform the inter-annotator agreement assessment and present the results of different evaluations.

We used a corpus of newspapers from Mexico. The news were dated between 1st of January, 2014, and 31st of October of 2014. All the news were manually selected as being about finances. The total of news articles is 300. From this we randomly selected 10 news.

From those selected news, we split each one into sentences. By doing this, we obtained 151 different sentences.

Then, we asked two annotators about the same age (21 yo), same gender, and same level of studies (last year of bachelor in language and arts), to label each sentence in each article as positive, negative or neutral, with no further explanation. The annotators did not know that another annotator was doing the same task. Finally we compared the sentences that both annotators labelled and compare if they agreed in the selected category or not. With those labels we created a confusion matrix containing in columns the distribution of categories given by annotator 1 and in rows those given by annotator 2.

After doing this we obtained the following results.

For the positive category, we got:

Table 2. Sentences labelled as positive

		Annotator 1		
		Yes	No	Total
Annotator 2	Yes	**38**	7	45
	No	45	**61**	106
	Total	83	68	**151**

As it can be seen in table 2, annotator 1 and annotator 2 agreed in 99 sentences. From this, they agreed that 38 were positive and 61 were not positive, thus they could be neutral or negative. It is also observable that while Annotator 1 labelled a total of 83 sentences as positive and 68 as not positive, Annotator 2 thought that just 45 sentences were positive, while 106 were not positive. So it is feasible to think that the agreement is as low as $\frac{38+65}{151} = .6556$. However, this is just the observed agreement.

For this particular case we have that $P(e) = .4799$. So in order to know the real agreement between the two annotators, it is needed to subtract the probability of chance agreement from the observed agreement. For this particular case we know that the agreement without chance is $\kappa = 0.3378$. This would be interpreted as having a **fair** strength of agreement, according with table 1.

However, if we compute the maximum kappa for this category, we get that $\kappa_{max} = 0.5161$. so, if we take this into account, the proportional kappa, normalised by its maximum, would be of $\kappa = 0.6545$. So in this case we could see it as having a **substantial** strength of agreement.

When we compared the sentence labelled as neutral, we found the distribution expressed on table 3.

Table 3. Sentences labelled as neutral

		Annotator 1		
		Yes	No	Total
Annotator 2	Yes	**34**	46	80
	No	8	**63**	71
	Total	42	109	**151**

In table 3 we observe that agreement is similar to that observed in table 2, that is, both annotators agreed that there were 34 neutral sentences and 63 non-neutral sentences. However, when looking at the totals, it is noticeable that while Annotator 2 thought there were 80 neutral sentences in total, Annotator 1 thought there were just 42. This mean that Annotator 2 thought there were 71 sentences that were not neutral, thus either positive or negative, while Annotator 1 thought there were 109 sentences being not neutral.

Because of this, the observed agreement is similar to, but a little lower than, that observed in positive. This is $P(o) = 0.6424$. Also the probability of chance agreement is a little higher than that for positive. For the neutral class, we computed $P(e) = 0.4868$. So, when we take into account the two facts, a little lower $P(o)$ and a little higher $P(e)$, it is normal to expect that κ coefficient would go lower, as it is the case. In this particular class, we computed a $\kappa = 0.3032$. This means that according to the interpretation table, we would evaluate this as having a **fair** strength of agreement.

As expected, the maximum kappa for this category is similar to that computed for positive. In this particular case we have that $\kappa_{max} = 0.5097$. So taking this into account, we get that the proportional kappa for the maximum expected agreement would be $\kappa = 0.5949$. This is considered to have a **moderate** strength of agreement between the annotators.

After looking at the sentences labelled as negative, we observed an increased agreement between the annotators. The table 4 shows how they distributed the sentences.

It is noticeable that agreement is much higher in this case. It can be observed, also, that differently from the previous cases, in this one, the matrix is symmetric. In this case, the annotators strongly agreed on what was not a negative sentence.

Table 4. Sentences labelled as negative

		Annotator 1		
		Yes	No	Total
Annotator 2	Yes	**15**	11	26
	No	11	**114**	125
	Total	26	125	**151**

Also both annotators agreed that there was just 26 negative sentences. This left 125 as non-negative sentences.

Thus, the observed agreement was $P(o) = 0.8543$. This is higher than the positive and neutral. This is because both annotators labelled many sentences as being non-negative.

We computed the hypothetical agreement by chance as being $P(e) = 0.7149$. This is also higher than in the other two cases. When these two factors are taken into account, the agreement coefficient is $\kappa = 0.4889$. This is notoriously higher than in the other two cases.

This means that it there will be more agreement on classifying sentences as negative or non-negative than in positive or non-positive and in neutral or non-neutral.

This is, annotator agreed on classifying 114 sentences as non-negative, and they both classified 125 as non-negative. However they didn't agreed much on classifying those sentences as positive or neutral. While annotator 1 considered 83 as being positive and 42 as being neutral (table 2, table 3), annotator 2 distributed their non-negatives almost flipped: 45 as being positive and 80 as being neutral.

For this particular case, the maximum possible kappa is of $\kappa_{max} = 1$. This is because the table is completely symmetrical.

This is interesting because it shows that it is much harder to agree on positiveness or neutrality of a message than it is to agree on their negativity.

In table 5 we can observe a summary of the results of the three tables.

Table 5. Kappa by category and it's interpretation

Category	*Observed Kappa*	*Interpretation*	*Maximum Kappa*	*Proportional*	*Interpretation*
Positive	0.3378	fair	0.5161	0.6545	substantial
Neutral	0.3032	fair	0.5097	0.5949	moderate
Negative	0.4889	moderate	1	0.4889	moderate

When we take into account the total distribution of the labelled sentences, we observe more clearly how the annotators classified each sentence. The next table summarises this process.

We can observe in a more detailed fashion how the non-class were distributed by each of the annotators.

Table 6. Sentence distributed by the three classes

		Annotator 1			
		Pos	Neu	Neg	Total
Annotator 2	Pos	**38**	6	1	45
	Neu	36	**34**	10	80
	Neg	9	2	**15**	26
	Total	83	42	26	**151**

As we discussed earlier when talking about the positive labels, both annotators agreed on 38 sentences being positive. However, annotator 1 also marked other 45 sentences as positive. From those other 45 sentences, annotator 2 thought 36 were neutral and 9 negative. On the other hand, annotator 2 just added 7 more sentences than the 38 they agreed as positive. From these sentences, annotator 1 labelled 6 as neutral and 1 as negative. In total, annotator 1 labelled 83 sentences as positive while annotator 2 just labelled 45.

For the neutral case, we find a similar scenario, nevertheless it is flipped over. In this category, both annotators agreed on labelling 34 sentences as neutral. On the disagreed sentences, annotator 1 labelled just 8 more as neutral; from these, annotator 2 thought that 6 were positive and two negatives. Contrastingly, annotator 2 foresaw 46 more sentences as being neutral; annotator 1 considered 36 of these as being positive and 10 as being negative. In total, annotator 1 labelled 42 sentences as neutral while annotator 2 almost doubled it and labelled 80 as neutral.

At labelling the negative classes, we find a little different picture. In this case, as stated, both annotators agreed as labelling 15 as negative. Also both labelled, in total, 26 as negative, but from the disagreed 11 of the annotator 1, annotator 2 views 1 as positive and 10 as neutral. On the other hand, from those 11 that annotator 2 esteemed negative and disagreed with annotator 1, the latter classified 9 as positive and two as negative.

The observed agreement for the three classes together would be $P(o) = 0.5762$. This does not seem as high as one might expect. Even though the chance of agreement has not being subtracted from that number. This chance is $P(e) = 0.3408$. Therefore we might assess the Cohen's correlation coefficient being as $\kappa = 0.3570$. This is, according with Landis and Koch [9], **fair** agreement.

Nevertheless, it is important to take into account that kappa assumes its theoretical maximum value of 1 only when both observers distribute codes the same. So, when it is the case that corresponding row and column are not identical, we need to compute the maximum possible value and use that maximum to interpret in a better way the agreement assessed.

We get that maximum kappa is $\kappa_{max} = 0.6182$. This mean that if we use this maximum as a help to interpret the actual kappa obtained, we could think that the value is around $\kappa = 0.5775$. This falls, according to Landis and Koch, in the lower rank of a substantial or higher rank of moderate agreement. Another interpretation that is possible to make based on the k_{max}, is that annotators will never agree in a higher degree. This mean that best agreement would be a moderate agreement.

5 Discussion

In this section we will discuss the results obtained during the experiment and will try to interpret them.

Based on the observations made on this experiment, we might notice several things. The most noticeable is that, when labelling opinions in the news, annotators without specific training will not agree in a high degree, specially when deciding if a sentence is positive or neutral.

We also noticed that the peak of agreement was achieved when deciding whether a sentence is negative or not. For this class, both annotators agreed that there were few negative sentences. Another interesting observation, is that even though both annotators agreed on tagging 26 sentences as negative, there were some sentences classified as positive by one annotator and negative by the other. This proofs that people will, at least sometimes, strongly disagree on the polarity of a sentence.

Another worthy observation is that while one annotator had a strong inclination for labelling sentences as positive (83), the other thought that many sentences were neutral (80). If we take into account that annotators are the same gender, same age and has same level of education, we might think that these differences are explained for their own experiences. This means that people, therefore annotators, will think of a sentence as being positive or neutral (or even negative) depending on their own perspectives and expectations of how the newspaper text should deal with the topic that it is about.

6 Conclusion and future work

From this experiment we can conclude that any classification system should take into account that agreement among annotators is more or less small. Because of this, it is possible to think that, under these conditions, any opinion mining system developed with this kind of corpus, will always be biased. This might be either a model bias or a learning data bias (if the system is non-supervised) or by the annotators that were used for labelling training data (in the case of supervised learning systems).

As future work we think it is important to compare these results with more data and assess agreement with other inter-annotator agreement metrics. We also think that it would be interesting to compare if the agreement is more or less the same when changing the input data, this is, with films or product reviews, with political speech, etc. Another thing we want to do in a future, is to compare the results after giving clear instructions of what to consider negative, neutral and positive to the annotators.

Acknowledgements

This work was done with the support given by the project IN403016 DGAPA-PAPIIT. The authors also want to thank the Consejo Nacional de Ciencia y Tecnología (CoNaCyT) for the support given through grant 387405 and project CB2012/178248.

References

1. Agarwal, B., Mittal, N.: Machine learning approach for sentiment analysis. In: Prominent Feature Extraction for Sentiment Analysis, pp. 21–45. Springer (2016)
2. Bermingham, A., Smeaton, A.F.: A study of inter-annotator agreement for opinion retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 784–785. ACM (2009)
3. Brants, T.: Inter-annotator agreement for a german newspaper corpus. In: LREC (2000)
4. da Cunha, I., Torres-Moreno, J.M., Sierra, G.: On the development of the rst spanish treebank. In: Proceedings of the 5th Linguistic Annotation Workshop. pp. 1–10. Association for Computational Linguistics (2011)
5. Deng, L., Wiebe, J.: Mpqa 3.0: An entity/event-level sentiment corpus. In: Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies (2015)
6. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC. vol. 6, pp. 417–422. Citeseer (2006)
7. Kaewphan, S., Van Landeghem, S., Ohta, T., Van de Peer, Y., Ginter, F., Pyysalo, S.: Cell line name recognition in support of the identification of synthetic lethality in cancer from text. Bioinformatics 32(2), 276–282 (2016)
8. Lacson, R., Raja, A.S., Osterbur, D., Ip, I., Schneider, L., Bain, P., Mita, C., Whelan, J., Silveira, P., Dement, D., et al.: Assessing strength of evidence of appropriate use criteria for diagnostic imaging examinations. Journal of the American Medical Informatics Association p. ocv194 (2016)
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
10. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining text data, pp. 415–463. Springer (2012)
11. Molina, A., SanJuan, E., Torres-Moreno, J.M.: A turing test to evaluate a complex summarization task. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 75–80. Springer (2013)
12. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
13. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval. pp. 557–566. ACM (2010)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)

15. Rangel, I.D., Guerra, S.S., Sidorov, G.: Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomazein 29(1), 31–46 (2014)
16. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing. pp. 254–263. Association for Computational Linguistics (2008)
17. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)
18. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language resources and evaluation 39(2-3), 165–210 (2005)
19. Yong, C., Foo, S.K.: A case study on inter-annotator agreement for word sense disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99) (1999)
20. Zhang, L., Liu, B.: Aspect and entity extraction for opinion mining. In: Data mining and knowledge discovery for big data, pp. 1–40. Springer (2014)