Examining Independence in Ensemble Sentiment Analysis: A Study on the Limits of Large Language Models Using the Condorcet Jury Theorem

Baptiste Lefort[a,b,*,1], **Eric Benhamou**[a,c,1], **Jean-Jacques Ohana**[a], **Beatrice Guez**[a], **David Saltiel**[a] **and Thomas Jacquot**[a]

[a]Ai for Alpha
[b]CentraleSupelec
[c]Dauphine - PSL

Abstract. This paper explores the application of the Condorcet Jury theorem to the domain of sentiment analysis, specifically examining the performance of various large language models (LLMs) compared to simpler natural language processing (NLP) models. The theorem posits that a majority vote classifier should enhance predictive accuracy, provided that individual classifiers' decisions are independent. Our empirical study tests this theoretical framework by implementing a majority vote mechanism across different models, including advanced LLMs such as ChatGPT 4. Contrary to expectations, the results reveal only marginal improvements in performance when incorporating larger models, suggesting a lack of independence among them. This finding aligns with the hypothesis that despite their complexity, LLMs do not significantly outperform simpler models in reasoning tasks within sentiment analysis, showing the practical limits of model independence in the context of advanced NLP tasks.

1 Introduction

The integration of Natural Language Processing (NLP) into the financial sector has been pivotal in providing insights from textual data, evolving significantly with the release of Large Language models and generative models [4, 12]. In particular, NLP has allowed to do sentiment analysis. The terminology sentiments analysis emphasizes that the method aims at extracting market sentiments or tones and then at offering predictive insights thanks to its analysis [42, 38].

Compared to other NLP applications, financial sentiment analysis faces unique challenges [43]. Financial narratives often involve complex, domain-specific terminologies and exhibit a multiplicity of sentiments tied to different entities, which can render general sentiment analysis tools ineffective [28, 34, 39]. The complexity is further exacerbated by the nuances of financial news, which may be reactive rather than predictive, and the difficulty of integrating sentiment scores into practical investment strategies [46, 15].

Historically, sentiments analysis relied on simple machine learning techniques and predefined word lists [30]. Later on, with the introduction of more sophisticated NLP techniques, and in particular the deployment of models such as BERT and FinBERT, sentiment analysis has significantly refined its accuracy [11, 26]. With the emergence of Large Language Models (LLMs) like GPT, and particularly its variants tailored for conversational applications such as ChatGPT, a new paradigm in AI has been set forth [9, 35, 33]. These models offer enhanced capabilities in understanding and generating human-like text, presenting new opportunities to improve sentiment analysis within financial contexts.

However, it is not clear how these new models differ from previous ones, motivating for a detailed and theoretically sound comparison with previous ones. In this paper, we focus on examining the independence of ensemble sentiment analysis conducted through LLMs, utilizing the Condorcet Theorem to investigate the limits and capabilities of these models in financial applications. The goal is to determine the extent to which these models can operate independently and to what degree their analysis can be considered reliable and unique in the context of complex sentiment tasks. Understanding this independence is critical, as it directly impacts the efficacy of models in real-world financial decision-making scenarios.

The paper is structured as follows: Section 2 reviews related works. Section 3 outlines the primary contributions of our work. Section 4 presents the theoretical framework for the Condorcet Jury theorem for bagging on multiple class classifications. Section 5 presents a bagging experiment for doing sentiment analysis on various LLMs and highlights that there is no real improvement when using larger models. Section 6 discusses why the Condorcet Jury theorem does not hold in this setting, proving a lack of independence among these models and suggesting that more recent models do not perform better in reasoning tasks within sentiment analysis. Section 7 provides concluding remarks and future research directions.

2 Related works

The exploration of ensemble methods in natural language processing, specifically through the lens of the Condorcet theorem, has garnered considerable attention in recent literature. This section discusses works closely related to our research and explains our contribution or similarities.

1. **Ensemble Methods in Sentiment Analysis:** A significant portion of research in sentiment analysis has advocated for the use of

---

ensemble techniques to improve classification accuracy. For example, [18] demonstrated that ensemble classifiers could robustly enhance sentiment detection across various social media platforms by leveraging multiple model predictions. Their work primarily focused on conventional machine learning models and did not explore the application of large language models (LLMs) in their ensemble. Our study extends this area by incorporating LLMs like GPT 3.5, GPT-4 and a finetuned version of GPT3.5 into the ensemble framework, testing the limits and benefits of bagging such advanced models in sentiment analysis.

2. **Application of Condorcet Theorem in Machine Learning:** The theoretical underpinning of our study is based on the Condorcet Jury theorem, historically used in majority voting [8, 6]. This theorem proves that a majority vote for a binary decision among independent classifiers leads to superior collective accuracy under certain assumptions [31]. [41] used the Condorcet Jury theorem to explain why their ensemble deep neural networks to screen for Covid-19 and pneumonia from radiograph images were demonstrating improved diagnostic outcomes. In our case, we extend the proof of the Condorcet Jury Theorem to multi-class problems by employing the assumption of an Independent, Well-Trained, and Uniformly Biased (IWTUB) set of classifiers toward the correct alternative. Additionally, we apply the Condorcet Jury Theorem in a novel context, shifting from an image recognition problem to evaluating the independence of decisions among various large language models (LLMs) and simpler natural language processing (NLP) models within a sentiment analysis task.

3. **Independence of Classifiers in Advanced NLP Models:** The assumption of independence among classifiers is critical in applying the Condorcet theorem effectively. [13] provide an overview of ensemble methods for binary classifiers in multi-class problems, highlighting challenges in classifier independence. They noted that in many practical applications classifiers often exhibit correlated errors, which can significantly diminish the efficacy of ensemble methods. Our findings are in line with these perspectives, revealing that the anticipated independence among various LLMs, including state-of-the-art models like ChatGPT 4, is not sufficiently pronounced to harness the full potential of the majority vote mechanism dictated by the Condorcet theorem.

4. **Performance of LLMs Versus Simpler Models:** Contrary to several prior studies that highlighted the superiority of LLMs in complex reasoning tasks [33, 27, 24], our empirical results indicate only marginal performance improvements when these models are included in an ensemble for sentiment analysis. [48, 23] also offer a critical perspective on the effectiveness of LLMs in sentiment analysis, questioning their practical superiority in sentiments analysis. Likewise, [47] observed experimentally that although LLMs consistently outperform Smaller Language Models (SLMs) in a few-shot learning context ([9]), they struggle on more complex models. This is consistent with our findings that large language models (LLMs) fail to outperform simpler models such as FinBERT [3] and DistilRoBERTa [37] in sentiment analysis tasks, suggesting their decisions are not independent.

3 Contributions

This paper seeks to enrich the field of natural language processing (NLP) by leveraging the Condorcet Jury Theorem within the specific context of sentiment analysis. Sentiment analysis requires sophisticated multi-class classification approaches. The key contributions can be summarized as follows:

1. **Theoretical Contribution:** We expand the Condorcet Jury Theorem, which to our knowledge has only been proven for binary decisions, to address the more complex scenario of multi-class classification. This extension relies on the new introduced concept of IWTUB set. Doing multi-class is crucial for sentiment analysis as sentiments typically involves categorizing texts into multiple classes beyond simple binary labels, such as positive, negative, and neutral or even potentially more granular sentiments.

2. **Empirical Validation of Model Non Independence:** Using a majority classifier with multiple NLP models, including fine-tuned FinBERT, DistilRoBERTa, GPT-3.5, and GPT-4, we found that majority voting did not improve performance. This suggests a lack of independence among the models, as indicated by the IWTUB set's last condition. Unlike traditional statistical tests like Pearson's chi-squared, Spearman's rank correlation, and mutual information, which assess dependencies between dataset variables, the Condorcet Jury Theorem uses voter competence and independence to predict the accuracy of group decisions, focusing on decision-making processes rather than dataset characteristics.

3. **Insights on LLMs:** Our results suggest a significant overlap in the decision-making processes for financial sentiment analysis of both compact and advanced models and a potential lack of reasoning for generative LLMs like GPT.

The paper's intuition is detailed in figure 1 for enhanced clarity on the overall methodological approach.

4 Extension of Condorcet jury theorem to multi-class

The Condorcet Jury theorem [10], was initially proven for a jury that should vote for a yes or no question. It states that if there is a majority preference for one option, then that option should be chosen as the overall winner in a collective decision-making process, provided that individual jury members are better at deciding than a blind guess and that there are independent [25, 32]. Mathematically, if we change individual jury members into machine learning classifiers and assume that each classifier follows the same law, an ensemble majority voting bagging classifier should perform better than any individual classifiers. In this section, we will see how to extend these results to a multi class classification. We will need to assume that classifiers are somehow similar. This is the core assumption that will be challenged in our LLMs experiment.

Definition 1. **IWTUB Set:** We say that a set of n classifiers $\{C_1, C_2, \ldots, C_n\}$ is an IWTUB (Independent, Well-Trained and Uniformly Biased towards the correct Alternative) Set if all the classifiers satisfy the following four conditions:

1. Identical Distribution: Classifiers have the same distribution.

2. Better Than Random: Each classifier C_i has an accuracy better than random guessing over the k possible predicted labels, i.e., for each C_i and for the true label t, the probability $P(C_i(x) = t)$ is more than $\frac{1}{k}$. For binary classification scenarios, this threshold stands at 0.5. In our specific context, where there are three possible classes (positive, indecisive or negative), the threshold adjusts to one-third.

3. Uniform Distribution Among Incorrect Alternatives: Each classifier C_i has the same probability which is lower than random guessing for choosing the incorrect alternatives, i.e., for each C_i and for any incorrect label $y \neq t$, the probability $P(C_i(x) = y)$ is less than $\frac{1}{k}$,
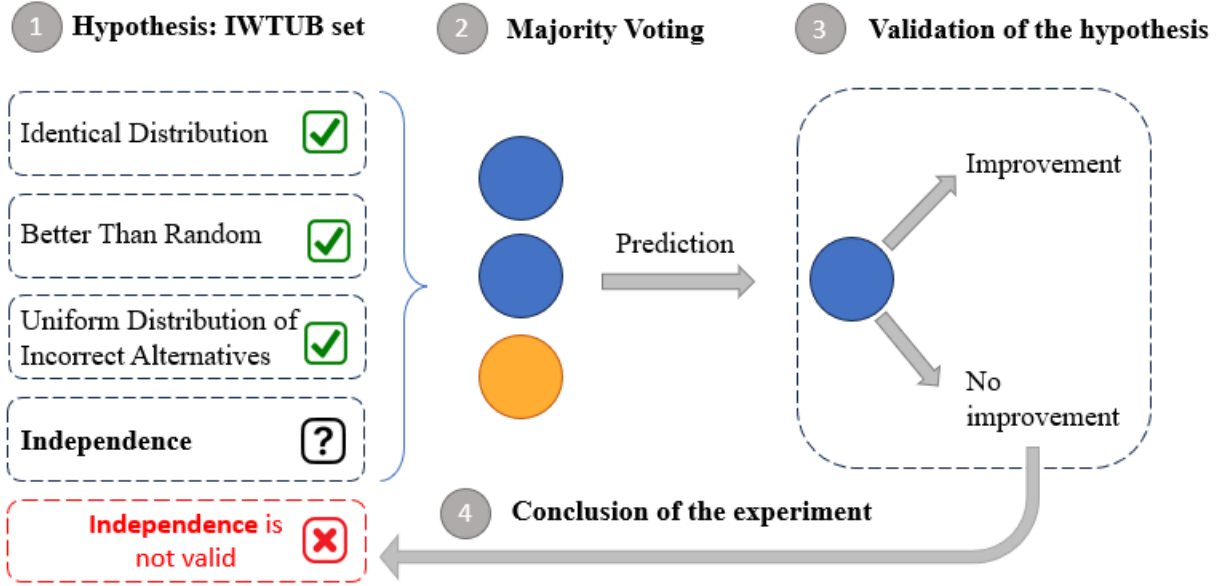
Figure 1. Procedure diagram of models independence validity revocation in the Condorcet's theorem.

4. Independence: Classifiers make their predictions independently of each other.

Among all the assumptions, the last one (independence of models) is usually quite problematic as there are often correlation in the way models handle data. Under the assumption of IWTUB Set of classifiers, we can extend the Condorcet theorem to multi class classifications problems as follows.

Theorem 1. **Condorcet Jury Theorem for Classifiers:** For an IW-TUB Set of classifiers $C_1, C_2, \ldots, C_n$, the majority vote classifier C_{bag} exhibits higher accuracy than any individual classifier within the ensemble. When the number of classifiers tends to infinity, the majority classifier converges in probability to the true label. Conversely, if base classifiers perform worse than random guessing, as the number of classifiers tends to infinity, majority classifier will diverge in probability from the true label and hence be worse than the random classifier.

Proof. The proof is an adaptation of the traditional two possible predicted labels presented for instance in [36]. It relies on two arguments. The Khintchine weak law of large numbers is used to establish the convergence in probability of the majority classifier applied to the set of our classifiers to the correct solution. Recall that a majority vote classifier is transitive [22, 21] meaning that changing the order of classifiers does not change the outcome. Recall also that a majority vote classifier is a fair function [16], meaning any true label plays same role and that applying a random permutation on the false labels does not change the result under the assumption of uniform distribution among incorrect alternative.

A classifier C_i takes some inputs X (in our NLP task some words) and provides an answer y (in our case a numerical sentiment) indicating which label is the true label among possible answers that are integers ranging from 1 to k. Because of fairness, there is a symmetry in true labels. Proving the result for one specific label will ensure that the result holds for any other true label. So without loss of generality, given some inputs X, we can assume that the true label is the largest label k to make our computation simpler.

In this specific case, a classifier would perform better than random guessing if its probability to spot the true label is higher than random guessing. Because of the symmetry of the problem, random guessing probability is $1/k$. Hence, if we define a the constant, that we also refer in the paper as the advantage of our classifier C_i over randomness, the probability to claim the true label is:

$$P(C_i = k) = \frac{1}{k} + a \tag{1}$$

with $a > 0$. Because our classifiers have equal probability of choosing the incorrect alternatives, the alternative probabilities for choosing any incorrect label $l \neq k$ is given by

$$P(C_i = l) = P(C_i \neq k) = \frac{1}{k} - \frac{a}{k-1} \tag{2}$$

In other words, we will choose the incorrect label with a probability dminished from the random guess by the fraction of the advantage over the wrong labels in number $k - 1$.

As alternative probabilities are positive, the advantage should satisfy $a \leq \frac{k-1}{k}$. Because of the assumption of independence and identical distribution of all our classifiers, each classifier has the same expectation which can be easily computed as the sum of possible outcomes times their corresponding probabilities, that can be split between the true label and the wrong labels.

$$\mathbb{E}[C_i] = kP(C_i = k) + \sum_{j=1}^{k-1} jP(C_i = j) \tag{3}$$

$$= \frac{k+1}{2} + a\frac{k}{2} \tag{4}$$

The result obtained above is very interesting. It says that the expectation of a classifier is equal to the average random guess among all possible labels, (the term $\frac{k+1}{2}$) plus the advantage part (the term $a\frac{k}{2}$).

If we impose that classifiers give equal weights to wrong label when making a wrong prediction, a series of classifiers will have a majority vote result equal to k if within the series of classifiers, the dominating class is the k class. If we denote by $N_a(\{C_1, ..., C_n\})$ (or simply N_a) the number of occurrence of a in the predicted labels of the series of classifiers $\{C_1, ..., C_n\}$, our assumption of uniform distribution among error will imply that assuming the wrong label j is equivalent to assuming the wrong label 1

$$N_j(\{C_1, ..., C_n\}) = N_1(\{C_1, ..., C_n\}) \quad \text{for } j \neq k \quad (5)$$

Hence, the dominating class will be k if and only if $N_k(\{C_1, ..., C_n\}) > N_j(\{C_1, ..., C_n\})$ for $j \neq k$ or simply $N_k > N_1$. We can notice that we can directly infer from the sum of predicted labels whether the dominating class is k or not as a classifier will have value of k N_k times and otherwise all other labels N_1 times. This gives us:

$$\sum_{i=1}^{n} C_i = kN_k + \frac{(k-1)k}{2}N_1 \quad (6)$$

$$= \frac{k(k+1)}{2}N_1 + k(N_k - N_1) \quad (7)$$

Hence,

$$N_k > N_1 \quad \Leftrightarrow \quad \sum_{i=1}^{n} C_i > \frac{k(k+1)}{2}N_1 \quad (8)$$

$$\Leftrightarrow \quad \frac{\sum_{i=1}^{n} C_i}{n} > \frac{k+1}{2} \quad (9)$$

where in the last part, we have used that $kN_1 + (N_k - N_1) = n$. In other terms, we get

$$\mathbb{P}\left[\text{Majority is correct}\right] = \mathbb{P}\left[\frac{\sum_{i=1}^{n} C_i}{n} > \frac{k+1}{2}\right] \quad (10)$$

The Khintchine weak law of large numbers [19] states that the empirical average $\frac{\sum_{i=1}^{n} C_i}{n}$ converges in probability to the expectation of the identically distributed classifiers. In other words, for any $\varepsilon_1, \varepsilon_2 > 0$, there exists N large enough such that for every $n \geq N$, $\mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} C_i}{n} - \mathbb{E}[C_i]\right| < \varepsilon_1\right) \geq 1 - \varepsilon_2$.

In particular for $\varepsilon_1 = a\frac{k}{2}$, we have for every $n \geq N$

$$\mathbb{P}\left[\text{Majority is Correct}\right] \quad (11)$$

$$= \mathbb{P}\left(\frac{\sum_{i=1}^{n} C_i}{n} > \frac{k+1}{2}\right) \quad (12)$$

$$= \mathbb{P}\left(\frac{\sum_{i=1}^{n} C_i}{n} - \frac{k+1}{2} - a\frac{k}{2} > -a\frac{k}{2}\right) \quad (13)$$

$$\geq \mathbb{P}\left(a\frac{k}{2} > \frac{\sum_{i=1}^{n} C_i}{n} - \frac{k+1}{2} - a\frac{k}{2} > -a\frac{k}{2}\right) \quad (14)$$

$$\geq \mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} C_i}{n} - \mathbb{E}[C_i]\right| < \varepsilon_1\right) \quad (15)$$

$$\geq 1 - \varepsilon_2 \quad (16)$$

where in equation (15), we have use equation (4) to introduce the expectation of any classifier $\mathbb{E}[C_i]$.

The last equation states that for any $\varepsilon_2 > 0$, there exists N such that for every $n \geq N$, we have

$$\mathbb{P}\left[\text{Majority is Correct}\right] \geq 1 - \varepsilon_2 \quad (17)$$

As equation (17) holds for any ε_2, this effectively proves that the probability of taking the true label would converge to 1. Hence, we have effectively proved that majority vote function would choose the true label with a probability that converges to 1 as the number of classifiers becomes large and tends to infinity. Last but not least, it is worth noticing that we have used in our proof that the advantage a is constant among all classifiers in equation (15).

A few remarks apply. The result can generalised to non constant advantages functions and this is the subject of theorem 3. The Khintchine weak law of large numbers is justified as the variance of the average of predicted values of our classifiers is finite. We could have also be tempted to use the Kolmogorov strong law of large numbers to establish that the convergence of the average of predicted values of our classifiers converges almost surely to the expected value of our i.i.d. classifiers [20]. However, in pure generality, since the convergence is not necessarily uniform on the set where it holds, it does not imply that with probability 1, we have that for any $\varepsilon > 0$, the inequality $\left|\frac{\sum_{i=1}^{n} C_i}{n} - \mathbb{E}[C_i]\right| < \varepsilon$ holds for all large enough n, which would be problematic. As a matter of fact, it turns out that the latter does hold as again the variance of variance of the average of predicted values of our classifiers is finite, which somehow circles us back to to the Khintchine weak law of large numbers. □

Corollary 2. Additionally if the independence assumption does not hold, the majority classifier C_{bag} should not perform better than the best classifier.

Proof. straightforward consequence of theorem 1 □

4.1 Non Uniform Classifiers

Up to this point, our analysis has assumed that all classifiers are uniformly distributed and equally accurate. We will now relax these assumptions to consider classifiers that vary in performance but are still more effective than random guessing. To explore how these non-uniform classifiers compare to random chance, we refine the concept of *Classifier Advantage*, which measures the extent to which a classifier outperforms random guessing.

Definition 2. **Classifier Advantage:** Let t be the true label. For a Uniformly Biased Towards the Correct Alternative (UBTCA) classifier C_i, we define its positive advantage a_i as its advantage or edge over random guessing:

$$a_i = P(C_i = t) - \frac{1}{k} \quad (18)$$

Remark. To have a probability above random guessing, the advantage a_i should be positive. Because our classifier has equal probability of choosing the incorrect alternatives, the alternative probability for $l \neq t$ is given by $P(C_i = l) = P(C_i \neq t) = \frac{1}{k} - \frac{a}{k-1}$. This leads in particular to an upper bound for the advantage to ensure that probabilities are positive: $a \leq \frac{k-1}{k}$.

We can now prove the condition of these classifiers relative to random guessing, establishing that the sum of advantages should grows significantly faster than $\sqrt{n}$, or equivalently that the advantage term should converge to zero slower than the inverse of the square root of the number of classifiers.

Theorem 3. **Condorcet Jury Theorem for Non-Equal Classifiers:** A majority vote classifier for non-equal classifiers, biased towards

the correct alternative, will converge to the true solution in probability as their number increases, provided the partial sum of advantages grows faster than the square root of their count:

$$\frac{\sum_{i=1}^{n} a_i}{\sqrt{n}} \underset{n\to\infty}{\to} \infty. \tag{19}$$

Proof. Consider a sequence of independent and identically distributed (i.i.d.) random variables $C_1, C_2, \ldots$, which we refer to as classifiers. Each classifier C_i can produce multiple potential outcomes, but we assume a uniform distribution among errors. Using the symmetry of our problem, without loss of generality we can assume that the true label among k possible ones is the largest value to make calculations easier. The assumption of equal distribution among errors transforms the initial multinomial setting into a binomial case, where the classifier either predicts the true label $t = k$ or the wrong labels $j \neq k$.

In a balanced binomial scenario, the probability of success (correct classification) and failure (incorrect classification) is traditionally set to $1/2$. However, given our setup where classifiers do not distinguish between the types of incorrect labels and the error is uniformly distributed across k possible outcomes, the probability of an correct classification for each trial is $1/k$. Denoting as in traditional statistical textbooks the partial sum $S_n = \sum_{i=1}^{n} C_i$ of the outcomes of our classifiers C_i, the sequence $\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right)$, converges in distribution to a normal distribution $\mathcal{N}(0, \sigma^2)$ as $n \to \infty$, where σ^2 is the variance of our i.i.d.s classifiers

In our specific settings, classifiers are not any more identically distributed as their advantage function a_i is not constant. However, we can still apply the Lyapunov central limit theorem (see [7] p. 371), or the central limit theorem for triangular arrays (see [17] Theorem 5.15) to see that the central limit theorem still holds as the individual probabilities of true classification $p_i = \frac{1}{k} + a_i$ converges or equivalently that the advantage functions a_i converges and consequently that the individual variance of our classifier converges to a finite value which we denote Σ and we will determine later.
Hence, in our case, we have

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right) \xrightarrow[n\to\infty]{} \mathcal{N}(0, \Sigma) \tag{20}$$

Let us refine our computation. For each classifier, we can compute their expectation and variance as follows

$$\mathbb{E}[C_i] = kP(C_i = k) + \sum_{j=1}^{k-1} jP(C_i = j) \tag{21}$$

$$= \frac{k+1}{2} + a_i\frac{k}{2} \tag{22}$$

$$\mathbb{E}[C_i^2] = k^2P(C_i = k) + \sum_{j=1}^{k-1} j^2P(C_i = j) \tag{23}$$

$$= \frac{(k+1)(2k+1)}{6} + a_i\frac{k(8k-1)}{6} \tag{24}$$

$$\mathbb{V}[C_i] = \mathbb{E}[C_i^2] - (\mathbb{E}[C_i])^2 \tag{25}$$

$$= \frac{k^2-1}{12} + a_i\frac{k(5k-4)}{6} - a_i^2\frac{k^2}{4} \tag{26}$$

The latter equation providing individual variance $\mathbb{V}[C_i]$ can be seen as a quadratic function of a_i. It attains its maximum when its derivative given by $\frac{k(5k-4)}{6} - 2a_i\frac{k^2}{4}$ is null. This occurs when $a_i = \frac{5k-4}{3k}$.

In particular, for an advantage function a_i ranging from 0 to $\frac{k-1}{k}$, the individual variance $\mathbb{V}[C_i]$ is bounded. In addition, if we assume that our advantage function a_i converges to a finite value a_∞ as i tends to infinity, we have that the individual variance would also converge to a finite quantity Σ:

$$\mathbb{V}[C_i] \xrightarrow[i\to\infty]{} \frac{k^2-1}{12} + a_\infty\frac{k(5k-4)}{6} - a_\infty^2\frac{k^2}{4} \tag{27}$$

Hence the asymptotic variance Σ is given by:

$$\Sigma = \frac{k^2-1}{12} + a_\infty\frac{k(5k-4)}{6} - a_\infty^2\frac{k^2}{4} \tag{28}$$

Using the linearity of the expectation, we have

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \frac{\sum_{i=1}^{n}\mathbb{E}[C_i]}{n} = \frac{k+1}{2} + \frac{k}{2n}\sum_{i=1}^{n} a_i \tag{29}$$

$$\tag{30}$$

We have now reached a point where we can establish the proof of our theorem. The majority function will accurately indicate the true label if and only if:

$$\frac{S_n}{n} > \frac{k+1}{2} \Leftrightarrow \sqrt{n}\left(\frac{S_n}{n} - \frac{k+1}{2}\right) > 0 \tag{31}$$

$$\Leftrightarrow \sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right) + \frac{k}{2\sqrt{n}}\sum_{i=1}^{n} a_i > 0 \tag{32}$$

Using the asymptotic distribution provided in (20), and for large value of n, this will be equivalent to

$$\mathcal{N}(0, \Sigma) + \frac{k}{2\sqrt{n}}\sum_{i=1}^{n} a_i > 0 \tag{33}$$

Hence, should the following series diverge,

$$\frac{\sum_{i=1}^{n} a_i}{\sqrt{n}} \underset{n\to\infty}{\to} \infty, \tag{34}$$

thanks to dominated convergence, the inequality (33) will hold with a probability that converges to 1 as n tends to infinity as:

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{N}(0, \Sigma) + \frac{k}{2\sqrt{n}}\sum_{i=1}^{n} a_i > 0\right] \tag{35}$$

$$= \mathbb{P}\left[\mathcal{N}(0, \Sigma) > -\lim_{n\to\infty}\frac{k}{2\sqrt{n}}\sum_{i=1}^{n} a_i\right] \tag{36}$$

$$= 1 \tag{37}$$

This concludes the first proof stating that the divergence of the series ensures the correctness of the majority with probability converging to one with the growing number of classifiers n. Conversely, should the ratio of the sum of advantages to the square root of their total number approach zero as n approaches infinity, the probability that the majority vote function identifies the correct outcome will converge to one half, given the symmetry of the asymptotic normal distribution $\mathcal{N}(0, \Sigma)$ around the origin. In other terms, the majority vote function will find the correct result like a random guessing, which concludes the proof. $\square$

Example 4. **Proper advantage functions** We can provide multiple examples of advantage series meeting the divergence condition given by condition (19).

- Logically, the simplest one is a constant strictly positive advantage: $a_i = \lambda > 0$. This is easy to check as

$$\frac{\sum_{i=1}^n a_i}{\sqrt{n}} \sim \frac{\lambda n \sqrt{n}}{2} \underset{n \to \infty}{\to} \infty \tag{38}$$

- Moreover, the advantage can have a logarithmic decrease to zero: $a_i = \frac{1}{\log i}$. Indeed, using equivalent, we can check that the series properly diverges to infinity as follows:

$$\frac{\sum_{i=1}^n a_i}{\sqrt{n}} \sim \frac{\sqrt{n}}{\log n} \underset{n \to \infty}{\to} \infty \tag{39}$$

- In the case of a polynomial decrease, it should be lower than the square root inverse as follows: $a_i = \frac{1}{i^\alpha}$ with $\alpha < 1/2$. For instance $a_i = \frac{1}{\sqrt[3]{i}}$ will work. In the general case, we can use the Riemann condition ([2]) to check the divergence as

$$\frac{\sum_{i=1}^n a_i}{\sqrt{n}} \sim n^{1/2-\alpha} \underset{n \to \infty}{\to} \infty \tag{40}$$

5 Bagging experiment

In this section, we introduce the results of the experiments made for comparing the state-of-the-art LLMs in financial sentiment classification: FinBERT [3], DistilRoBERTa [37] and GPT-4. We detail the framework and the individual abilities of each models. This experimental section enables us to validate the non-applicability of the Condorcet's theorem in this context and subsequently to question the assumption of models independence.

5.1 Experimental framework

We use a proprietary dataset of headlines generated from Bloomberg news called market wraps that span from 2010 to 2024 and is known to be of high quality as it is scrutinized by market professionals [1]. These headlines are a consolidated summary of daily financial news, done by human journalists specialized in finance, highly regarded and extensively followed by professionals within the financial sector. The database has about 65,000 rows representing on average 15 major headlines per day over 15 years. For each headlines, the database comes with the corresponding next day returns of major equities markets, hence providing a way to systematically provides a news sentiment that is predictive of future impact by definition. We can therefore validate if a model is able to predict effectively a predictive news sentiment. This dataset is containing high-quality data and is closely followed by the market players [14]. The labels are balanced as detailed in table 1. All these arguments make this dataset valuable for fine-tuning and accurately assessing the performance of the considered LLMs on the financial classification task.

Table 1. Number of instance in each class with percentage

Negative	Neutral	Positive
19254 (*31%*)	16202 (*27%*)	25795 (*42%*)

The dataset we present, as detailed in Table 2, is specifically designed to provide insights into economic events through news headlines, each linked to precise dates and associated sentiment classifications. This arrangement is crucial for analyzing how sentiments expressed in financial headlines correlate with market events and conditions, offering a robust foundation for temporal trend analysis over

an extensive period from 2010 to 2024. This allows for longitudinal studies examining the evolution of sentiment in response to global economic, political, and social changes.

Each row of the dataset contains a sentiment label—Positive, Negative, or Indecisive—systematically assigned by an algorithm, enhancing its utility for financial forecasting. The robustness of this dataset, especially compared to public datasets such as Financial PhraseBank [29] and StockNet Dataset [45], stems from its focus on actionable investment decisions without the noise often introduced by social media content or human annotation, which can be less predictive or non systematic and subjective. In particular, the direct linkage of each sentiment classification to subsequent market performance in this private dataset ensures that the labels reflect predictive market events, as discussed in the literature [40, 5, 44].

The primary objective of classifying sentiment in financial headlines is to generate predictive signals for the market. In our analysis, if a Large Language Model (LLM) can accurately classify the sentiment of all stocks mentioned in the headlines, it would yield a valuable trading signal. This capability underscores the dataset's effectiveness in enhancing market outcome predictions.

Table 2. Dataset Sample

Date	Headline	Sentiment
2010-01-04	Oil Prices Surge Above $81 a Barrel Due to U.S. Cold Weather	Positive
. . .	. . .	. . .
2014-07-02	Euro and Pound Gain on Hawkish Central Bank Policies	Negative
. . .	. . .	. . .
2020-11-06	S&P 500 Index Sees First Retreat in a Week	Indecisive
. . .	. . .	. . .
2024-01-30	General Motors Beats Q4 Expectations, Expects Profit Growth	Positive

5.2 Models assessment

At this step, we examine the individual performances of the state-of-the-art models. They have both different architecture and different number of parameters. However they are considered as comparable for this classification task [3, 33].

Table 3. Model Performances Before Fine-Tuning

Model	F-score ↓	Precision	Recall
GPT-4	**0.47**	**0.48**	**0.49**
GPT-3.5	0.45	0.47	0.47
DistilRoBERTa	0.44	0.45	0.45
FinBERT	0.44	0.46	0.46
Random Classifier	0.33	0.33	0.33

Table 3 illustrates that generative models such as those in the GPT series do not outperform encoder-only models based on the BERT architecture, despite GPT's substantially larger number of parameters. This finding challenges the assumed predictive superiority of GPT models over more compact architectures like BERT. However, it is evident that all models significantly surpass the random benchmark, indicating their capacity to learn from the labeled data provided.

The extensive Bloomberg dataset utilized in this study facilitates precise fine-tuning of these models. We divided the dataset into two segments: one for training the models and the other for evaluating their performance, as shown in Table 4, which reports the models' results on the test set. It is noteworthy that at the time of writing this paper, GPT-4 could not be fine-tuned, but we were able to fine-tune the latest version of GPT-3.5. Given the similar performance of the non-fine-tuned models, we anticipate comparable results when GPT-4 becomes available for fine-tuning.
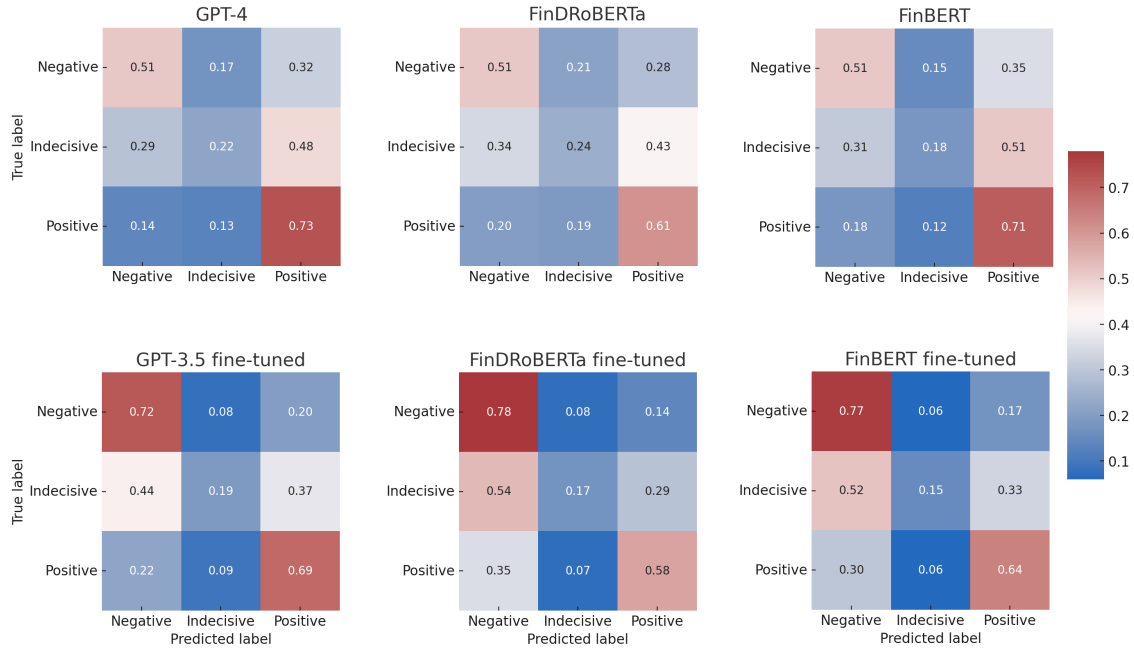
Figure 2. Confusion matrices for each assessed models

The fine-tuned models demonstrate no significant differences in performance among themselves; all exhibit uniform improvements. This indicates that the fine-tuning process equally enhances each model, without any single model demonstrating distinct superiority. Thus, while individual models are equally enhanced, none emerges as distinctly superior.

Table 4. Fine-Tuned Model Performances

Model	F-score ↓	Precision	Recall
SFT GPT-3.5	**0.51**	**0.54**	**0.53**
SFT DistilRoBERTa	0.49	0.53	0.51
SFT FinBERT	0.50	0.54	0.52
Random Classifier	0.33	0.33	0.33

Based on these performances, we apply an ensemble strategy. Since all the models have comparable performances we could expect a performance improvement if the Condorcet's hypothesis hold. Below is the list of the two different bagging configurations, including Supervised Fine-Tuned (SFT) models.

- Bagging 1: SFT GPT-3.5 + SFT DistilRoBERTa + SFT FinBERT
- Bagging 2: All the models SFT and No SFT

Table 5. Ensemble method performance compared with the best solo performer model SFT GPT-3.5. The difference of metric between the Bagging methods and the SFT GPT-3.5 model is provided.

Model	△ F-score ↓	△ Precision	△ Recall
Bagging 2 vs SFT GPT-3.5	0.01	0.00	-0.01
Bagging 1 vs SFT GPT-3.5	0.00	0.01	-0.01

Table 5 provides that no significant improvement is made with a bagging method. The ensemble classifier could not improve the overall performance.

6 Discussion

Our experiments have confirmed that three of the four critical hypotheses necessary for applying Condorcet's theorem hold, namely.

1. **Identical Distribution** The confusion matrices in Figure 2 demonstrate that the classification performance distributions across the models are consistent. Both fine-tuned and non-fine-tuned models exhibit comparable results across different classes, with the probability distributions of each LLM being markedly similar.

2. **Better than random** As evidenced by the confusion matrices in Figure 2 and performance data in Tables 3 and 4, all LLMs in our study perform significantly better than a random classifier.

3. **Uniform distribution among incorrect alternatives** The confusion matrices in Figure 2 indicate that the error distribution among incorrect choices is uniform across models. The similarity in error distribution and the high correlation in the instances where models fail further support this hypothesis.

Since the three hypothesis hold and are confirmed by the experiments, the last condition of IWTUB set can not hold, namely the independence hypothesis of the LLMs models used in our bagging experiment. In particular, this indicates that GPT models cannot be considered as independent or different models from simpler models like FinBERT and DistilRoBERTa. In particular, this suggests that the reasoning capacity of GPT models may not be very different form the smaller models and that they lack the capacity to reason on complex financial tasks like financial sentiment analysis.

7 Conclusion

This paper extends the Condorcet Jury Theorem, previously limited to binary decisions, to multi-class classification scenarios. This advancement incorporates the new concept of IWTUB set, introduced to facilitate the extension of the theorem to multi-class classification systems. Our empirical investigation of the applicability of a majority vote classifier revealed that despite the advanced capabilities of LLMs, such as GPT-3.5 or 4, there are only marginal enhancements in predictive accuracy when combined with smaller models. This outcome suggests a significant overlap in the decision-making processes of these models, underscoring their limited independence, which undermines the effectiveness of the majority voting mechanism from the Condorcet Jury Theorem and highlights the lack of reasoning abilities of LLMs for complex NLP task like sentiment analysis. The use of the Condorcet Jury theorem is very complementary to traditional statistical tests for independence, such as Pearson's chi-squared test, Spearman's rank correlation, and mutual information, that only focus on establishing the absence of a relationship between two variables in a dataset as the Condorcet Jury theorem leverages the assumption of voter competence and independence to predict the likelihood of correct decision-making in group settings. Future research should complement this approach by investigating other methods, like statistical ones to assess the non-independence among classifiers and exploring scenarios where large language models (LLMs) are effectively implemented in ensemble settings.

References

[1] F. S. Alzazah and X. Cheng. Recent advances in stock market prediction using text mining: A survey. *E-Business-Higher Education and Intelligence Applications*, pages 1–34, 2020.

[2] T. M. Apostol. *Mathematical analysis*. Addison-Wesley, 1974.

[3] D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[4] D. W. Arner, J. Barberis, and R. P. Buckley. The evolution of fintech: A new post-crisis paradigm. *Geo. J. Int'l L.*, 47:1271, 2015.

[5] M. N. Ashtiani and B. Raahemi. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, 217:119509, 2023.

[6] D. Austen-Smith and J. S. Banks. Information aggregation, rationality, and the condorcet jury theorem. *American political science review*, 90(1):34–45, 1996.

[7] P. Billingsley. *Probability and Measure*. Wiley, 2 edition, 1986.

[8] P. J. Boland. Majority systems and the condorcet jury theorem. *Journal of the Royal Statistical Society Series D: The Statistician*, 38(3):181–189, 1989.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] M. de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, 1785.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, XX(XX):XX, 2018.

[12] G. Fatouros, G. Makridis, D. Kotios, J. Soldatos, M. Filippakis, and D. Kyriazis. Deepvar: a framework for portfolio risk assessment leveraging probabilistic deep neural networks. *Digital finance*, 5(1):29–56, 2023.

[13] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776, 2011.

[14] S. García-Méndez, F. de Arriba-Pérez, A. Barros-Vila, and F. J. González-Castaño. Detection of temporality at discourse level on financial news by combining natural language processing and machine learning. *Expert Systems with Applications*, 197:116648, 2022.

[15] I.-B. Iordache, A. S. Uban, C. Stoean, and L. P. Dinu. Investigating the relationship between romanian financial news and closing prices from the bucharest stock exchange. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5130–5136, LREC, 2022. LREC. URL http://www.lrec-conf.org/proceedings/lrec2022/index.html.

[16] G. James, T. Hastie, and C. Sugar. Majority vote classifiers: theory and applications. *Machine learning*, 36(1):105–139, 2001.

[17] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2 edition, 2002.

[18] J. Kazmaier and J. H. Van Vuuren. The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*, 187:115819, 2022.

[19] A. Y. Khintchine. Sur la loi des grands nombres. *Comptes rendus hebdomadaires des s'eances de l'Acad'emie des sciences*, 188:477–479, 1929.

[20] A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, 1933.

[21] L. I. Kuncheva and J. J. Rodr'ıguez. A majority vote classifier with probabilistic guarantees. In *2014 22nd International Conference on Pattern Recognition*, pages 1478–1483. IEEE, 2014.

[22] L. I. Kuncheva and C. J. Whitaker. Classifier ensembles for changing environments. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2006.

[23] B. Lefort, E. Benhamou, J.-J. Ohana, B. Guez, D. Saltiel, and D. Challet. When small wins big: Classification tasks where compact models outperform original gpt-4. *Available at SSRN 4780454*, 2024.

[24] B. Lefort, E. Benhamou, J.-J. Ohana, D. Saltiel, B. Guez, and D. Challet. Can chatgpt compute trustworthy sentiment scores from bloomberg market wraps? *arXiv preprint arXiv:2401.05447*, 2024.

[25] C. List and R. E. Goodin. Epistemic democracy: Generalizing the condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001.

[26] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, ICLR, 2021. ICLR.

[27] A. Lopez-Lira and Y. Tang. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Electronic Journal*, XXX(XX-XX):XX, 4 2023. ISSN 1556-5068. doi: 10.2139/SSRN.4412788. URL https://papers.ssrn.com/abstract=4412788.

[28] T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

[29] P. Malo, A. Sinha, P. Takala, O. Ahlgren, and I. Lappalainen. financial_phrasebank. *Hugging Face*, 2014. doi: 10.5281/zenodo.4723759. URL {https://huggingface.co/datasets/financial_phrasebank}.

[30] Y. Mejova. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*, page 5, 2009.

[31] H. Mercier and O. Morin. Majority rules: how good are we at aggregating convergent opinions? *Evolutionary Human Sciences*, 1:e6, 2019.

[32] M. Morreau. Democracy without enlightenment: A jury theorem for evaluative voting. *Journal of Political Philosophy*, 29(2):188–210, 2021.

[33] OpenAI. Gpt-4 technical report, 2023.

[34] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

[35] K. I. Roumeliotis and N. D. Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

[36] Á. R. Sancho. On the probability of the condorcet jury theorem or the miracle of aggregation. *Mathematical Social Sciences*, 119:41–55, 2022.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[38] R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[39] M. Sensoy, L. M. Kaplan, and et al. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 3183–3193, 2018.

[40] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

[41] G. Srivastava, N. Pradhan, and Y. Saini. Ensemble of deep neural networks based on condorcet's jury theorem for screening covid-19 and pneumonia from radiograph images. *Computers in Biology and Medicine*, 149:105979, 2022.

[42] P. C. Tetlock. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168, 6 2007. ISSN 1540-6261. doi: 10.1111/J.1540-6261.2007.01232.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2007. 01232.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261. 2007.01232.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261. 2007.01232.x.

[43] M. Wankhade, A. C. S. Rao, and C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[44] F. Z. Xing, E. Cambria, and R. E. Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[45] Y. Xu and S. Cohen. Stocknet dataset. *Papers With Code*, 2018. doi: 10.5281/zenodo.1161185. URL {https://paperswithcode.com/dataset/stocknet-1}.

[46] C. Yuan, Y. Liu, R. Yin, J. Zhang, Q. Zhu, R. Mao, and R. Xu. Target-based sentiment annotation in chinese financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5040–5045, LREC, 2020. LREC. URL http://www.lrec-conf.org/proceedings/lrec2020/index.html.

[47] B. Zhang, H. Yang, and X.-Y. Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*, 2023.

[48] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. Sentiment analysis in the era of large language models: A reality check, 2023.